Scalable Data Analytics,
Scalable Algorithms, Software Frameworks
and Visualization ICT-2013 4.2.a

Project	**FP6-619435/SPEEDD**
Deliverable	**D7.4**
Distribution	**Public**

http://speedd-project.eu

Final Evaluation Report of SPEEDD Dashboards for Credit Card Fraud management

Chris Baber, Natan Morar, Faye McCabe, Sandra Starke,
(University of Birmingham)

Ivo Correira
(Feedzai)

Status: FINAL

December 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	D7.4
Deliverable Title	Final Evaluation Report of SPEEDD Dashboards for Credit Card Fraud management
Contractual date of delivery	M36 (January 2017)
Actual date of delivery	December 2016
Relevant Task(s)	WP7/Tasks 7.4
Partner Responsible	Feedzai
Other contributors	UoB
Number of pages	27
Author(s)	C. Baber, N. Morar, F. McCabe, S. Starke, and I. Correira,
Internal Reviewers	A. Artikis
Status & version	final
Keywords	Evaluation, User Interface Design, Human Factors

Contents

Figures

Tables

Executive Summary

This deliverable reports the final evaluation of the dashboard design for the SPEEDD Credit Card Fraud use case. The initial prototype, described in D7.1, sought to reflect features that were common in user interface designs in the financial sector. This design was implemented in the first SPEEDD prototype dashboard. Following initial evaluation with analysts and laboratory trials (reported in D7.1 and D5.2) the user interface design was revised (D5.3). An initial report of the evaluation, focusing on usability, was presented in D7.3. This report considers how the dashboard designs have an impact on user performance for simulated fraud analysis activity.

The results from this study suggest that dashboard design, computer confidence and type of fraud had an impact on the manner in which participants approached the fraud analysis task.

In terms of dashboard design, decision times were much faster with the Detailed (2) dashboard than the Overview (1). One explanation for this difference in time is that participants opened more modals when using the Overview (1) than the Detailed (2) dashboard. Thus, the time to gather information was longer when using the Overview (1) because participants consulted more information sources.

In terms of computer confidence, in contrast to the study on Road Traffic Management (D8.6), there was no effect of computer confidence on decision time or number of modals opened. While participants were presented with the computer's confidence in the dashboards, this did not seem to influence their decision time or information search. However, there were differences between which decision participants made and the computer confidence, and a significant interaction between computer confidence and number of decisions made. This suggested that when the computer confidence was high, participants would be more likely to identify the transaction as a fraud, and when computer confidence was low, they would allow the transaction. This potentially points to the issue of trust and its impact on user decision making; participants would see a relatively high (although not perfect computer confidence) as sufficient to accept a recommendation of fraud.

In terms of type of fraud, there were differences in decision time but participants tended to open more modals for the Large Amounts fraud type than the others. This implies that there was some ambiguity and uncertainty in defining Large Amounts (in a transaction) as an index of fraud. Consulting the other sources of information could indicate that the participants were seeking to check and confirm the computer's recommendation.

Finally, a quantitative analysis was also performed, where good results for the precison and recall were obtained, as well as decreases in latency were observed as a result of the applied improvements in the complex event processing engine.

1. Introduction

1.1 History of the Document

Version	Date	Author	Change Description
0.1	20/11/2016	Chris Baber	First version of the document
0.2	14/12/2016	Chris Baber	Changes made
0.3	15/12/2016	Natan Morar	Internal review of report
0.4	31/12/2016	Ivo Correia	Adding results for precision, recall and latency

1.2 Purpose and Scope of Document

The purpose of this document is to report the final evaluation of the dashboards for SPEEDD in the Credit Card Fraud Management use case. In terms of evaluation, the aim is to show how the prototype should be used by the fraud operators.

1.3 Relationship with Other Documents

This document is related to the following deliverables: 7.1 User Requirements, 7.2 Initial Evaluation Report, 7.3 Interim evaluation of user interface for credit card fraud use case; D5.1 Design of User Interface for SPEEDD Prototype, D5.2 Design of User Interface for SPEEDD Prototype (year 2), D5.3 Design of User Interface for SPEEDD Prototype (year 3).

1.4 Sources of Information

Information was gathered from the experienced fraud investigation personnel in one of Feedzai's client organisation and through experimental comparison of the dashboards designed for SPEEDD.

2. Defining a Baseline

The analysis of credit card fraud has become highly automated in recent years [1, 2, 3, 4, 5]. However, there remain situations in which a human analyst might be required to either contact a cardholder to check a transaction or to review the decisions made by automated systems. The output from automated analysis can present challenges to human interpretation (for instance, in terms of appreciating the underlying rules to which the automated systems operate or in terms of understanding why some data is given precedence over others). Thus, there is a growing trend to the use of interactive data visualisation to support analysts [6]. Interactive data visualization offers a range of claimed benefits for users: *"Visual analysis software allows us to not only represent data graphically, but to also interact with those visual representations to change the nature of the display, filter out what's not relevant, drill into lower levels of detail, and highlight subsets of data across multiple graphs simultaneously."* [7, p.4]. While this provides a compelling set of claims, the research evidence to support these claims is less well developed. In particular, this paper considers how people interact with displays that vary in the level of detail they provide and how this influences users' decisions about whether to 'drill into lower levels of detail'.

2.1 Dashboards, Situation Spaces and Decision Spaces

A common trend is for data information visualization displays to be designed on an analogy with the 'dashboard' of an automobile. *"When properly designed for effective visual communication, dashboards support a level of awareness—a picture of what's going on—that could never be stitched together from traditional reports. Unfortunately, most dashboard products and most of the vendors that develop and sell them, fail to take full advantage of data visualization's power."* [7, p.5]. In such displays, a subset of available information is presented to the decision maker to provide a situation space [8], i.e., a summary of key data that correspond to a particular situation, which will 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 [8]. 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. It is plausible to assume that the dashboard might contain only the information needed to adequately characterise the situation space, and that the experienced decision maker knows the options available in their decision space. For example, in the fraud analysis considered in this report, the decision space consists of {allow a transaction, query the transaction with the card-holder, block the transaction}. This decision space could be expanded, for instance, by including such options as launch an investigation, search for transaction patterns which might indicate organised crime, use the transaction to define a pattern for automated analysis etc. We assume that the decision space is known to the decision maker and, consequently, need not form part of the dashboard. However, there are other forms of decision making in which it might be profitable to

display the decision space. For example, [9] describes an approach which maps the information content of a dashboard to the business processes that it is intended to support, e.g., through the selection of specific 'case' properties that reconfigure the information presented.

2.2 Dealing with Automation Reliability

In the scenario that we developed, when automated analysis seeks human support, this would imply that the confidence or coverage of the automation has fallen outside defined thresholds. Consequently, the human analyst would be called upon when the automated system was uncertain as to the most appropriate decision to make. In such instances, the human would need to intervene as rapidly and efficiently as possible. This requires the analyst to understand the situation space and decision space in which the automation was operating. A dashboard could offer a high-level perspective on the situation space, with the opportunity to drill-down into lower levels of detail to expand the analyst's understanding of the situation space in order to determine an appropriate option in the decision space. While this might appear to be a simple process for the analyst, there is some evidence that the level of confidence of the automation can have a bearing on how this process is approached. If the decision from automation is associated with a high confidence rating, then the role of the human could be to confirm this decision. This can lead to problems of automation bias [10, 11, 12, 13, 14] or complacency [15, 16, 17, 18] in which the human merely accepts the automation's output and does not contribute to the decision making. Conversely, if the decision from is associated with a very low confidence rating, then the human might ignore entirely an output from the automation, even if it might be useful [19].

Recently, it has been suggested that acceptance of (and, by implication, trust in) automated decision support can be considered in terms of strategic conformance (i.e., when the automation problem-solving style matches that of individual human), in terms of the solution to a problem and the strategy used to arrive at that solution [13]. For the credit card fraud domain, the solution could be defined in terms of the decision to block or allow a transaction on a credit card and the strategy could relate the use of available information in informing this decision. In this respect, one might anticipate users of an automated decision support system to align their responses to that of the system, providing the information available supported the decision and, perhaps, providing the automated system was confidence in its recommendation.

2.2.1 Human Analysis of Credit Card Fraud

As noted in section 3.1.2, the vast majority of credit card transactions will be assessed automatically. This is especially the case for regular, high-volume credit card use: in this scenario, transactions are automatically scored. If a transaction is scored as too risky, the card is automatically blocked and the transaction is followed up with the customer, usually through a call-center affiliated with the bank. On the other hand, automated transaction analysis can be considered as a form of screening for human analysts. In these instances, a given transaction might not quite fit the profiles used by the automated system (or there might be a requirement for a small proportion of decisions to be checked by a human operator in order to maintain confidence in the automated system's performance). This may happen when the automated scoring system is tuned to new or unusual fraud types by specialist fraud analysts. In our scenario, a human operator will review transactions and make a decision, presenting a hybrid between the above two scenarios.

The credit card industry is understandably very protective of the approaches used in the analysis of credit card fraud. While we have benefitted from discussions with a number of fraud analysts operating in the UK, Europe and the US, the following description presents a high-level account of the type of decision making in which analysts engage. Some of the approaches that could be used were reviewed and discussed in D7.2 Initial Evaluation report.

Figure 1 implies a decision process for credit card fraud analysis. This is not intended to represent analysis conducted by any individual organisation but more a general description of how analysis is approached. It is useful to appreciate this process in order to understand how we have designed the experimental trials for this paper (see section 4.2.2). It is worth noting at this point that some of the aspects operate at an organisational, rather than individual, level. In terms of the System Output, there might be a 'risk score' which is based on risk probabilities that are defined in terms of an organisation's set of risk factors which, in turn, depend on the specific risk model that the organisation employs. The risk model will be tailored for specific types of client, region, transaction etc., but could include such measures as number of transactions for an account in a given time period, amount of money in a transaction, number of cash withdrawals at automated teller machines etc. These risk models inform the design and operation of algorithms used by the automated system.

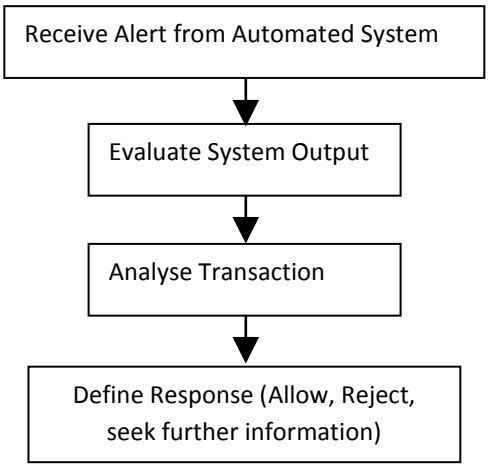

Figure 1: Decision Process for (human) credit card fraud analysis used in this paper

If a transaction meets the criteria that the algorithms apply, then it would be automatically blocked. However, if only some of the criteria were met or if there was some uncertainty concerning the criteria, the transaction might be presented to a human analyst. For the purposes of this paper, we are interested primarily in two main types of analyst involved in credit card fraud, each working with different levels of detail and at different levels of security.

Call-centre agents will perform customer verification on some suspicious transactions. In general, the decision to contact a customer would be made if the automated system was able to match some but not all of its criteria. In this instance, the call-centre agent would be presented with some of the transaction details together with the automated output (in the form of a score). In terms of the stages in figure 1, this activity would involve the top three boxes. The call-centre analyst would engage in

some form of sensemaking (combining information from the dashboard with information obtained from the cardholder) in order to define the situation space. In this instance, the 'situation space' could involve cases in which there is a legitimate and plausible explanation for the transaction, together with indications that this explanation might be questioned. The call-centre agent would then confirm that the transaction was acceptable or mark it as suspicious. This could involve reimbursing the card holder or could trigger further investigation. The Call-centre agent would follow a clearly defined script in order to establish whether a) the person at the other end of the phone is the genuine cardholder and b) whether the card holder made the purchase or whether it was made fraudulently. For this role, agents have access to transaction and customer details, both past and present. In third party fraud (card used by an unknown person), calling the customer is the only option to find out the true state of a transaction, at least. In contrast, first party fraud (malicious intent by card holder) or second party fraud (card user known to card owner) is more difficult to establish. For first and second party fraud, human intelligence is needed in order to determine whether answers to questions are genuine or fabricated.

In cases involving call-centre analysts, alerts are dealt with according to the risk level; alerts with the highest level of risk are worked on first (this is called 'priority mode'). The bank will set the score threshold so that the number of alerts to be dealt with per day can be matched by the number of employed staff. Banks may vary between 250 and 1500 FTE staff, and they will also show variation in interaction style depending on the customer demographic. The total number of cases to be processed by an operator following referral due to a flagged transaction is around 200 per day. Assuming that the analyst works an 8 hour shift with a meal break and two shorter rest-breaks, this could give a period of working time of 7 hours (420 minutes), which would give a limit of around 2 minutes per transaction. Assuming that the bulk of this time would be spent in a combination of speaking with customers or searching for additional information, then the time to make a decision on a suspicious transaction could be quite small, say measurable in tens of seconds.

Many credit card companies are seeking to remove call-centre agents entirely from their processes because they see these 'in-flight referrals' as adding undue interruption to the transaction process activity. This role is sometimes replaced by computerised solutions, which call or text a customer automatically to confirm the validity of a purchase.

In addition to call-centre agents, financial institutions employ fraud analysts who look for more general or emerging patterns of fraudulent behaviour. This could involve examining batches of 'exception reports' generated by the automated system, where the decisions made need to be verified. This involves a deeper analysis of the customer activity and transaction than would be performed by the call-centre agent, e.g., in terms of checking a history of transactions by that cardholder (or using that card) and comparing this with a set of behaviours and data which can indicate probability of fraudulent behaviour. In terms of figure 1, most of the actions could be performed by this analyst, although their decisions would be influenced by the policy and strategy of the organisation.

For the purposes of this report, we assume that the dashboard will be used mainly by call-centre analysis (to prepare and guide conversation with the card holder) or their line managers (monitoring the distribution of calls and the processing rates).

3. Experimental Comparisons of Dashboards

3.1 Dashboard design for our fraud analysis task

Two dashboards were designed for the experiment presented in this paper. The basic concept of a dashboard for fraud analysis was explored in D7.1 User Requirements and Scenario Definitions, and D7.2 Initial Evaluation report provided examples of commercial products. As our design is intended to fit with the Feedzai platform, we wanted to keep the layout of at least one of the dashboards as close as possible to the house style of Feedzai.

The first dashboard (figure 2) is an Overview (1) dashboard. This dashboard is designed with the office manager in mind, who might be interested in the overall activity of card-users and the activity of a team of call-centre analysts under the manager's control. Instead of presenting fine-grained transaction information, it presents the user with a list of all the automation-flagged patterns (table in the "Event List" window at the left of the screen in figure 2) along with statistics computed for the geographical region selected on the map in the window in the right. These statistics are the number of transactions investigated, total transactions flagged by the automated system, average transaction amount and average transaction volume per for the selected region. Each of the bits of information is shown in a separate small coloured window at the top of the screen in figure 2. The "Event List" window shows the transaction number, the fraud pattern identified, the computer certainty associated with this flag and the current status of the pattern (fraud, contact, allow – in case a decision had been made by the analyst – and not investigated – in case a final decision is not yet present).

At the bottom of this window there are four buttons: three of them reflect the decision space for this task (as outlined in section A: {allow a transaction, query the transaction with the card-holder, block the transaction}) and the fourth button is labelled "Explain" which brings up more data related to the selected pattern (figure 3). The map window on the right also has an "Explain" button, and this brings up information related to the selected country (figure 4).

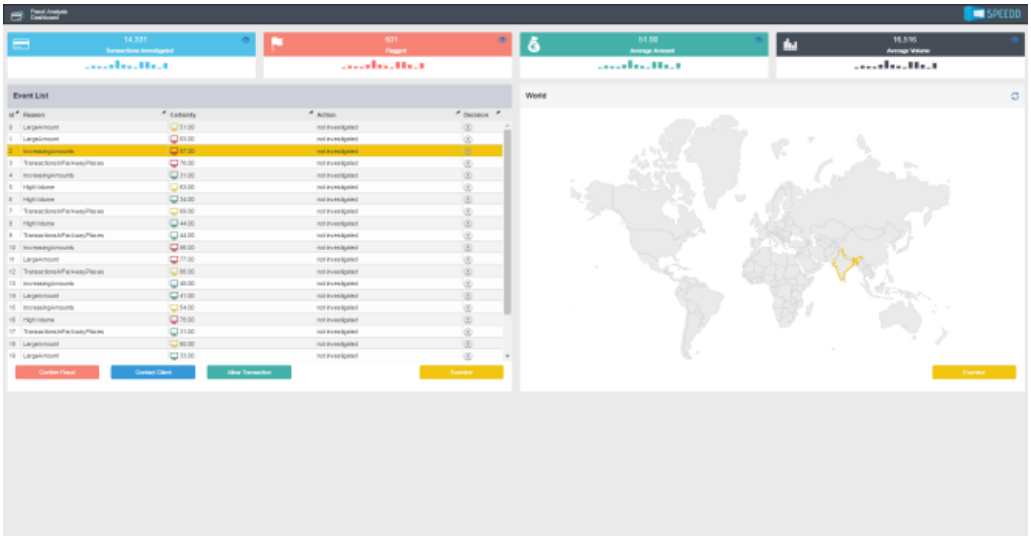

Figure 2: Dashboard 1: Overview (1)

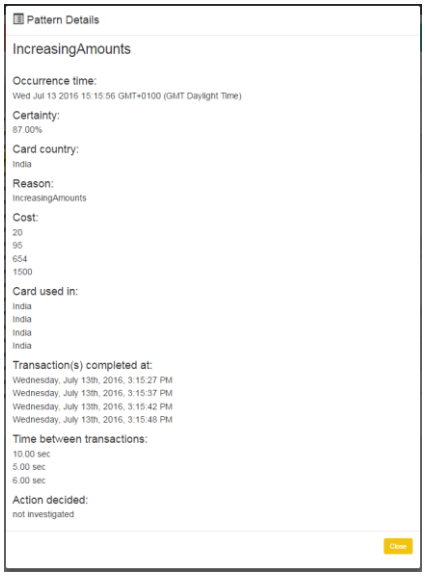

Figure 3: List modal

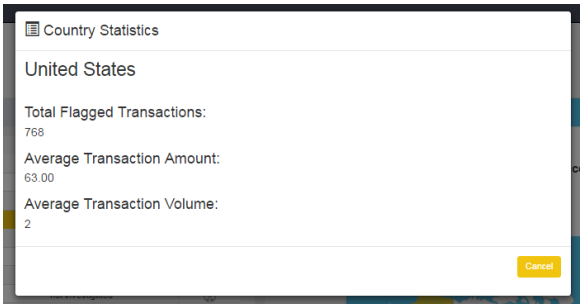

Figure 4: Country modal

The second dashboard (Detailed (2)) is designed to support the low-level decision-making of call-centre analysts. In contrast to the interface presented previously, this one (shown in figure 5) consists of lower-level information which directly relates to flagged transactions and provides no Overview (1) statistics. The window on the left labelled “Patterns to Investigate” is the same as the “Event List” window in the Overview (1) dashboard. However, the window on the right hand side contains information directly relevant to the flagged transaction type. This window changes its content depending on the fraud type selected in the “Patterns to Investigate” window. At the top of the “Pattern View” window there are two visualisations. The first one illustrates the transaction sequence which triggered the automated flagging. On this visualisation both the number of transactions and the interval between them is shown. The second visualisation (on the right) relates to the transaction

amounts. If only one transaction is present then this is presented as a single bar, whereas in case of a sequence multiple bars are shown in chronological order. Below these visualisations, a map is shown where the country or countries in which the transaction/s took place is/are highlighted. Finally, at the bottom of the window, the date and time of flagging is shown along with the computer certainty and reason for flagging (the transaction flag name). Similar to the Overview (1) dashboard, "Explain" buttons are present in both windows and serve the same function as those in figure 2.

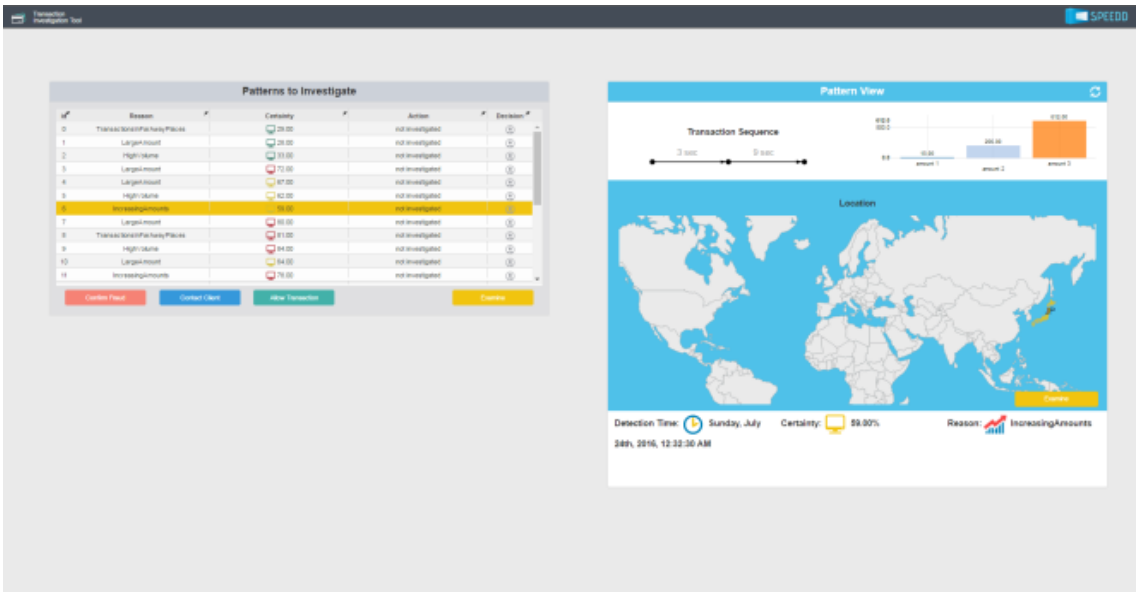

Figure 5: Dashboard 2: Detailed (2)

3.2 Method

The experiment was designed to explore three questions:

i. does information exploration activity (defined in terms of decision time or drill-down) differ when people are presented with Overview (1) or Detailed (2) dashboards?
ii. does information exploration activity vary with computer confidence?
iii. do users alter their information exploration activity in response to different task demands (i.e., different fraud types in this study)?

Each suspicious transaction had an associated automation confidence to inform the analyst as to a probable type of fraud.

3.2.1 Participants

27 people took part in the experiment [17: male; 10: female; age range: 22-29]. None of the participants had experience of working in the credit card industry or financial sector. Given the nature of the experiment and the observation that fraud schema are highly company-specific, we felt that it was appropriate to recruit such a sample because they would respond to the information according to their organisational policy and strategy, which as noted previously will vary across organisations. Having said this, section 5 presents the results of this study performed by a small number of experienced credit fraud analysts in a partner company of Feedzai. For our recruitment of participants in study one, we assume that people educated to degree level and given training on the fraud patterns

to identify would provide a reasonable proxy with call-centre agents, whose role is primarily to follow the script provided to them (as outlined in section 3.1.3). As we were presenting information that reflected both the fraud patterns and data used in the SPEEDD project, it was felt that a cohort of trained participants would be appropriate. Therefore, each participant was first trained to criterion (see below) before the experiment began. In this way, we would have a homogenous user group from which to explore the impact of dashboard designs on decision performance.

3.2.2 Procedure

The study was approved by the University of Birmingham ethics committee. All data were anonymised and participants provided informed consent. Following a briefing on the task and training in using one of the user interfaces (allocated on appearance), participants completed a series of practice tasks. This provided an opportunity for them to familiarise themselves with the user interface and also allowed the experimenters to ensure that each participant had reached an acceptable level of proficiency before beginning the trial. Participants were provided with an aide memoire which defined the four fraud patterns and these definitions were explained.

Participants were given a demonstration of how these patterns could be recognised from the data presented in a dashboard and the other windows and, following 2 or 3 familiarisation trials (to become accustomed to interacting with the user interface on a dashboard), were asked to process up to 10 examples. Once participants were able to correctly process 5 examples, training stopped and the main experiment began.

Each participant investigated 24 patterns with the dashboard. When this was completed, the process was repeated (i.e., familiarisation, training to criteria and experimental trials) with the other user interface.

The set of 24 patterns (for each dashboard) were randomised across all participants in order to minimise order effects and were defined in terms of four fraud patterns {increasing amounts; transactions in faraway places; large amount; high volume of transactions} and three levels of automation confidence {low (δ51%); medium (51% - 69%); high (ε70%)}. Each fraud pattern was presented twice under each confidence level.

The design of the experiment also considered the need to drill-down for further information and this was balanced across the two user interfaces. The participants needed to call up one of two modals (pop-up windows) shown in figures 9 and 10. However, use of modals is only relevant to some of the frauds (as shown in table 1) and so efficient performance could be defined in terms of modal use.

Table 1: need for drill-down

Fraud Type	Overview (1) Dashboard	Detailed (2) Dashboard
increasing amounts	Call up the list XXX (by clicking 'Examine' in the event list)	Use data on dashboard
transactions in faraway places	Call up the list XXX (by clicking 'Examine' in the event list)	Use data on dashboard

large amount	Use data on dashboard	Call up the country XXX (by clicking 'Examine' on a country)
high volume of transactions	Use data on dashboard	Call up the country XXX (by clicking 'Examine' on a country)

The Dependent Variables for the experiment (for the independent variables dashboard, fraud pattern and computer confidence) were: mean time to submit a decision; number of fraud patterns investigated further; efficiency score; subjective rating of workload.

The data were tested for normality and, where the data were normally distributed, analyzed using Analysis of Variance and pairwise comparison using t-tests. Where appropriate, the ANOVA is corrected using Greenhouse-Geisser corrections. Results were considered significant for $p < 0.05$. Effect sizes are reported using partial eta squared (η_p^2) and we assume, following [22] that scores >0.14 can be considered large and >0.06 can be considered medium. For non-parametric comparison, analysis was conducted using a Friedman test. All statistical tests were performed using IBM SPSS v13.

Measuring Workload. While there are many ways to measure the cognitive effort (workload) that people experience in performing mentally demanding tasks, a popular set of measures rely on participants providing subjective estimates of their workload. These measures can be surprisingly robust, sensitive to changes in demands and correlate well with physiological measures. One commonly used subjective workload measure is the NASA TLX (Task Load Index) [23]. This is a rating scale with six workload dimensions. It can be administered in either a computer or paper based format. We used the paper and pencil version of the test[1]. The rating scales are presented as questions that the participants score on a scale of 1 (low) to 20 (high). The questions relate to mental demand, physical demand, temporal demand, effort, performance and frustration (figure 6).

1

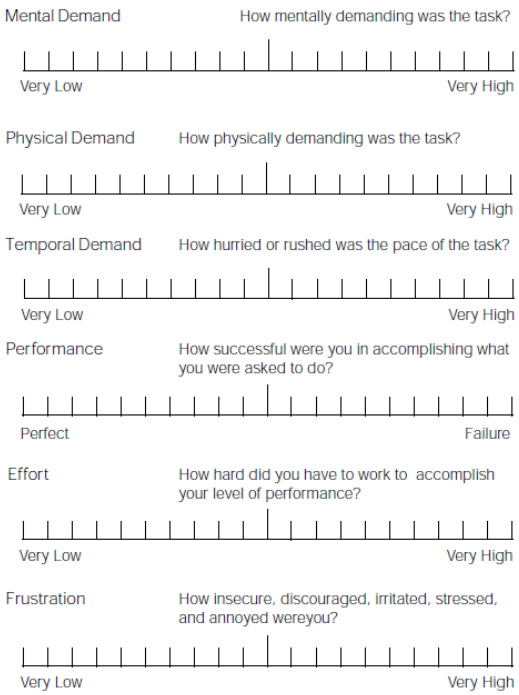

Figure 6: NASA TLX rating form

[http://humansystems.arc.nasa.gov/groups/tlx/paperpencil.html]

3.3 Results

Prior to analysis, data from two of the 27 participants were excluded because their average response times exceeded 2 standard deviations from the mean.

3.3.1 Impact of Dashboard design on decision time and information seeking

The mean decision times were normally distributed (Shapiro-Wilk p =.191 for Overview (1) and .690 for Detailed (2)). A paired samples t-test revealed significant difference in decision time between the two dashboard (t (24) = 3.136, p = 0.004). This difference is shown in figure 7.

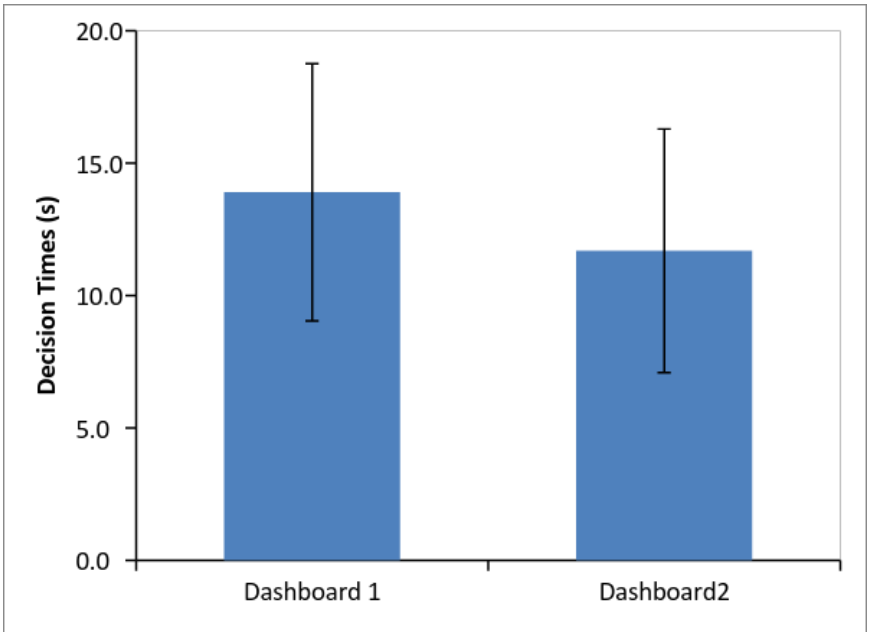

Figure 7: Comparison of decision times for the two dashboards

While there is a significant difference between the dashboards, visual inspection of figure 7 might lead one to conclude that the decision time for the two dashboards are more similar than the test result indicates. Plotting the difference in decision times(between using the Detailed (2) (1) and Overview (1) (2) dashboard), figure 8, shows that the difference is greater than 1, indicating that decisions were quicker when using the Detailed (2) dashboard.

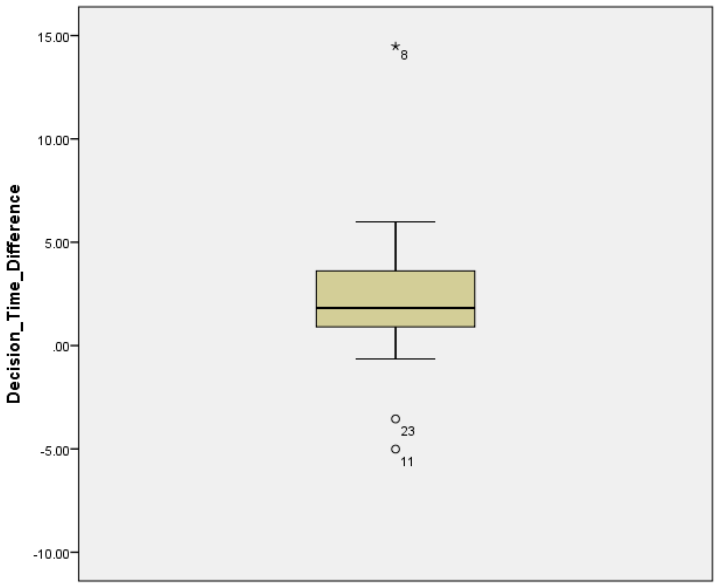

Figure 8: Differences in Decision Time between the two Dashboards

In terms of the average number of modals opened per decision for the two dashboards, while the Detailed (2) dashboard showed normal distribution in these data (Shapiro-Wilk, p = .871) the data for the Overview (1) were not normally distributed (Shapiro-Wilk, p =.000). Consequently, comparison was made using a Wilcoxon sign test. This showed that there was a significant difference in average number of modals opened (t = 4.088, p = 0.001). As figure 9 shows, participants tended to open more modals with the Overview (1) dashboard than with the Detailed (2), i.e., for every 4 modals opened in the Overview (1), they opened 1 modal in the Detailed (2) dashboard.

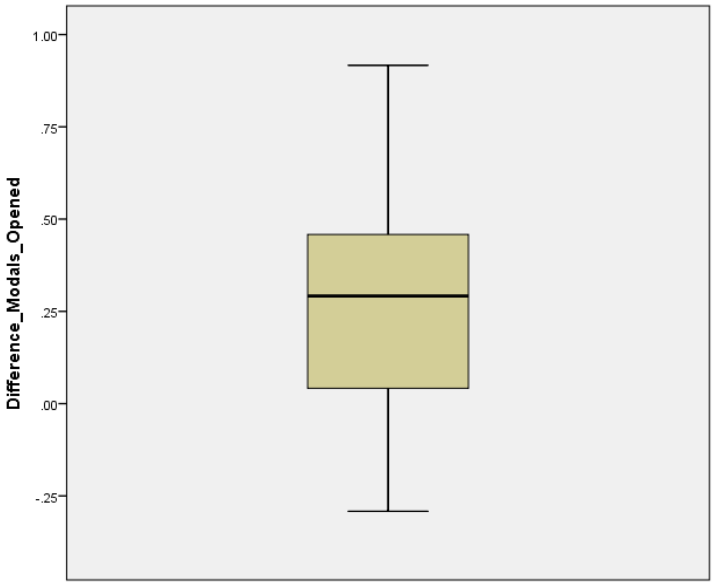

Figure 9: Difference between average number of modals opened with the two dashboards

3.3.2 Impact of Computer Confidence on decision time and information seeking

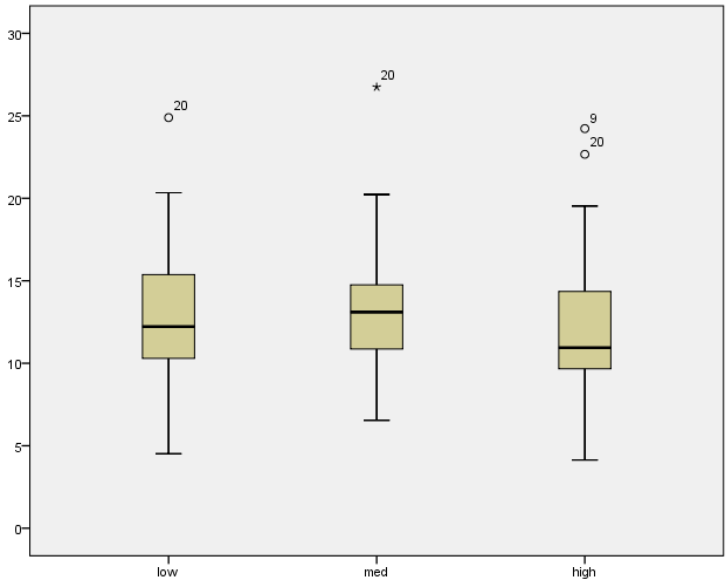

Figure 10: Mean decision time across computer confidence levels

Tests for normality indicated that all sets were normally distributed (Shapiro-Wilk = .645 for low; .111 for medium and .329 for high). A one-way ANOVA showed that there was no significant effect of computer confidence on decision time [F(2,74) = .357, p = .701]. As figure 10 illustrates, decision times are similar across confidence levels.

In terms of the number of modals opened, the data were normally distributed and so were analysed using ANOVA. There was no difference in number of modals opened across the three confidence levels. Participants responded to the level of computer confidence in terms of decision type. Figure 11 shows, the relationship between computer confidence and user decision and indicates that participants were more likely to allow a transaction when computer confidence was low (i.e., around 40%) and more likely to define a transaction as fraud when computer confidence was higher (i.e., around 65%). This produce a significant main effect for decision type [F (1,26) = 63.668 ; p = 0.0; η_p^2 = .71]. Furthermore, there was a significant main effect of user interface on this measure [F (1,26) = 5.341 ; p = 0.029; η_p^2 = .17]. There was also a significant interaction between user interface and decision type, although this was of medium effect [F (2,52) = 3.55 ; p = 0.047; η_p^2 = .12].

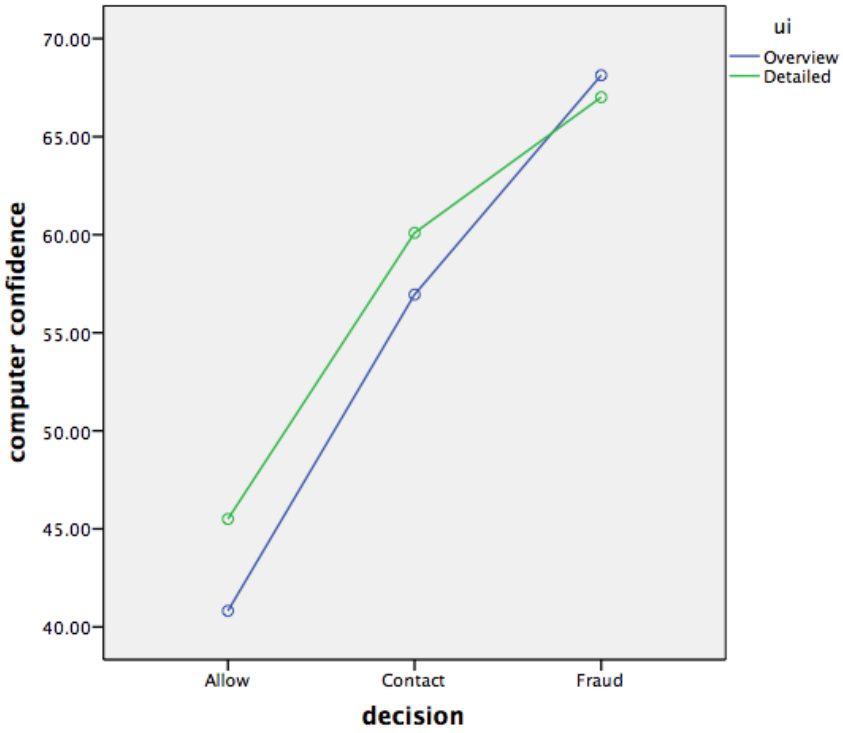

Figure 11: Relationship between computer confidence and user decision for the two dashboards

There was a significant main effect of computer confidence on the number of decisions made, of medium effect [F (1,26) = 3.250; p = 0.047; η_p^2 = .111]. There was a significant interaction between computer confidence and number of decisions (calculated using Greenhouse-Geisser correction) [F (4,104) = 24.451; p = 0.0; η_p^2 = .485].

3.3.3 Impact of fraud type on Drill-down activity

Comparing decision time in terms of fraud type, the data were normally distributed (Shapiro-Wilk: IA = .056; TF = .009; LA = .962; HV = .252). Thus, a one-way ANOVA was used and showed no significant effect of fraud type [F(3, 99) = .894, p=.447]. This is supported by figure 12.

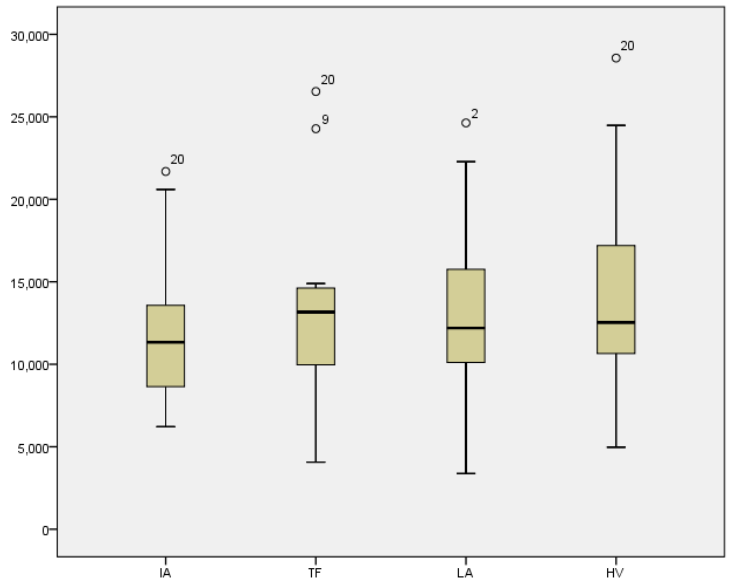

Figure 12: Decision times for each fraud type

Comparing the number of modals opened to investigate each fraud type, the data were not normally distributed and so a Friedman ANOVA was applied. This showed a significant effect of fraud type [x2 (3) = 13.455, p = .004]. This is supported by figure 13.

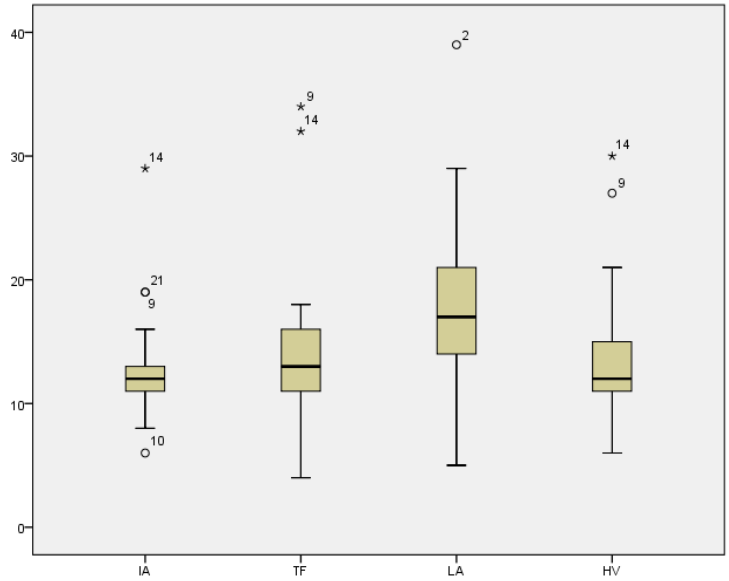

Figure 13: Number of modals opened for each fraud type

Pairwise comparisons show that the differences are between LA and the other fraud types, but not between pairs of the other fraud types. Thus, the difference between the fraud type LA (Large Amounts) and the others is sufficient to explain the main effect. It seems that participants consistently checked more information when confronted with the fraud type LA.

Table 2: Comparison of fraud types (* = p< 0.05, **p<0.001)

	IA	TF	LA	HV
IA	-	.183	.00**	.304
TF		-	.012*	.819
LA			-	.003*
HV				-

3.3.4 Workload

There was no difference in workload score between the two conditions (figure 13).

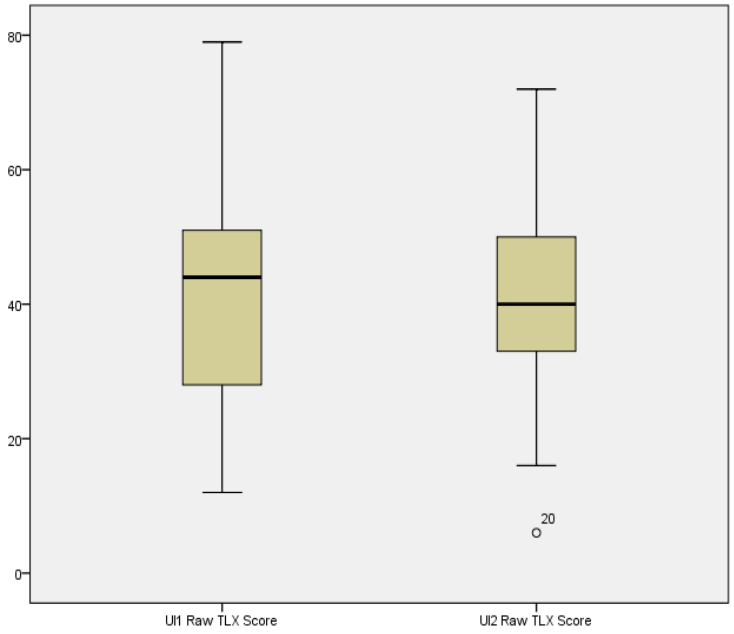

Figure 14: Comparison of Subjective Workload rating between conditions

3.4 Discussion

The experiment was designed to explore three questions.

The first question was, does decision making and information exploration activity differ when people are presented with Overview (1) or Detailed (2) dashboards? In terms of dashboard content, it is shown that the Detailed (2) dashboard has a faster decision time and leads to a lower level of drill-down than the Overview (1) dashboard. To some extent the differences in decision time indicate differences in strategy of information use and suggest that the availability of more information (in the Detailed (2) dashboard) improves decision time.

In terms of use of information, participants used more information sources, in both dashboards, for the 'large amount' fraud type. This could be indicative of participants seeking to acquire as much information as possible prior to making their decision when the fraud type might be less easy to define as anomalous. This could relate to the suggestion that people seek more information to increase their confidence in a decision, without an increase in decision accuracy [24, 25].

The second question was, does decision making vary with automation confidence? In terms of computer confidence, there is a clear and consistent impact on decision types. When the computer was confident, participants were likely to indicate that the transaction was a fraud. When the computer confidence was low, participants were more likely to allow the transaction. Interestingly, the decision to contact the cardholder tended to be similar across levels of computer confidence (with a modest drop at high confidence).

Relating these findings to the notion of strategic conformance [13], it can be seen that participants adapted their decision to align with the confidence of the computer and that their use of available information was, to some extent, dependent on the type of fraud that was being flagged. In this previous paper, there was the suggestion that "conformance may only be relevant for expert users who hold consistent and well-developed decision-making strategies" (p.50). However, for this study it is suggested that the clear presentation of the situation space (in terms of information related to credit card transactions) and the well-defined decision space (in terms of the relationship between possible decision and available information) was sufficient to support conformance.

Two versions of a dashboard for analysing credit card fraud were compared. Four types of credit card fraud were considered. Results showed that the Detailed (2) dashboard resulted in faster decision times and that use of drill-down (to find additional information) was lower for the Detailed (2) dashboard but also varied depending on the type of fraud being investigated.

4. Performance by Expert Analysts

The experiment was packaged for online delivery so that participants in a partner company of Feedzai's could complete it. The experiment was completed by 4 fraud analysts. While this number is too small to apply statistical analysis, it provides an indication of the how the dashboards might affect performance on the task and allows for comparison with the students who completed the experiment reported in chapter 4.

Figure 14 shows that the experts completed the tasks, between 8s and 27s, and there is a tendency for dashboard 2 to result in faster time than decision 1. This suggests a similar pattern to that observed in the students.

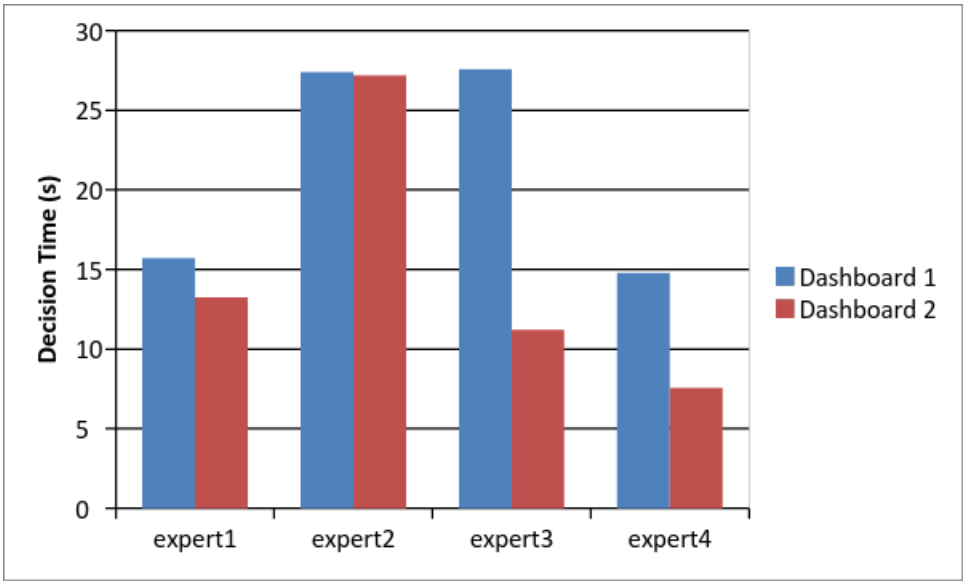

Figure 15: Average task completion time for 4 experts

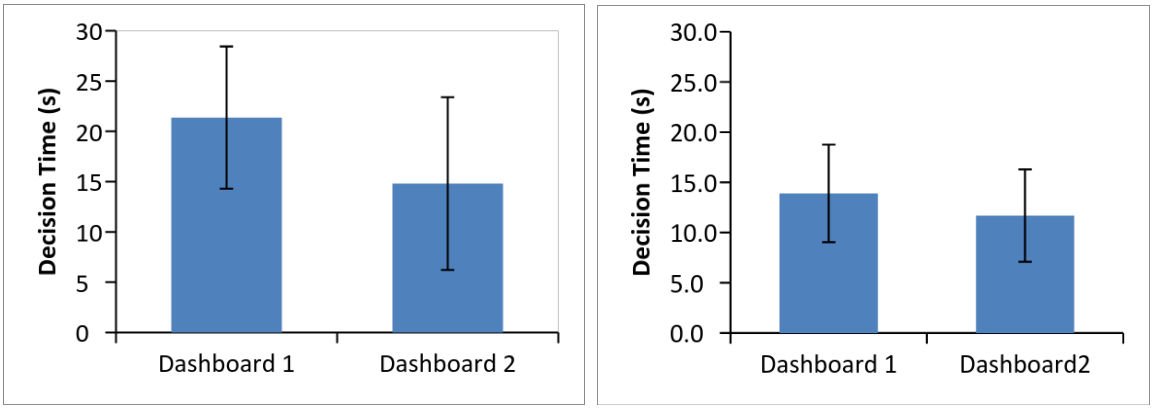

Figure 16: Comparison of average response times for experts and students

Comparing the average time of the experts with that of students (figure 15), it is clear that the experts tended to show more variability in the decision times (although this is likely to be due to the small sample size) and tended to be slower than the students. Having said this, the decision times for the experts tend to fall inside the distribution of those for the students and so one can assume that there is little difference in average time between these two groups.

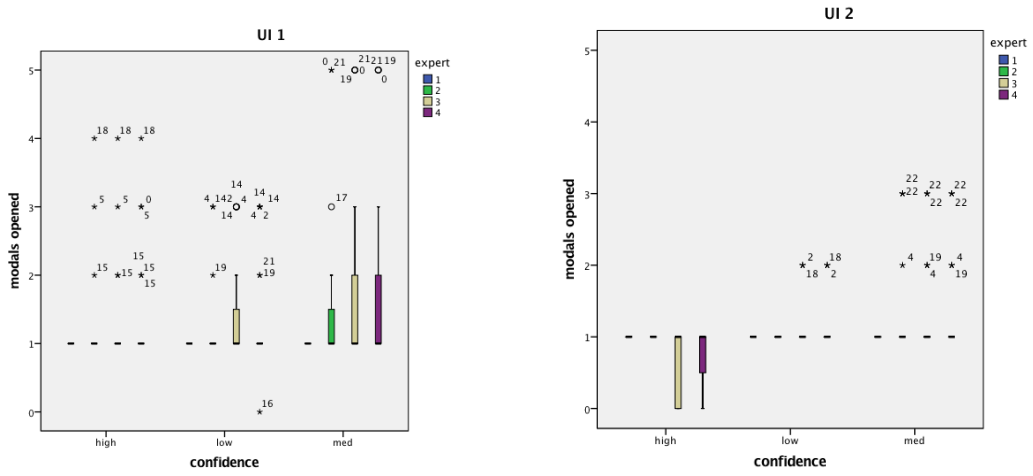

Figure 17: Modals opened in terms of confidence

Experts seem to open more modal windows for the medium and low confidence levels than for the high level (Figure 17). Moreover, while using UI 2, analysts tend to be using less extra information, in some situations they don't look for extra information at all. Linking this back to the experiment design, when using UI1, modals need to be brought up for Increasing Amounts and Transactions in Faraway Places. Conversely, when using UI2, extra information needs to be investigated for Big After Small and Flash Attack patterns in order to make an informed decision. However, what seems to happen is that fraud analysts tend to oversample, looking at extra information even when this is not required. Nevertheless, they oversample to a lesser degree when using UI 2 compared to UI 1.

Discussion. There is a difference between students and experts in terms of strategy. When we look at the decisions made, we find that the experts tend to select 'contact customer' first, then 'fraud' and then 'allow'. This was applied to all fraud types except for 'large amounts'. This suggests that, in line with the comments made by the fraud analysts we interviewed in D7.1, the 'gold standard' for determining whether a transaction is fraudulent is to contact the card holder. A further interesting finding is that experts' decisions to contact the customer and to flag as fraud (i.e., block card) remain relatively constant with computer confidence, however the decision to allow further transactions decreases as the computer confidence increases.

For the students, the order of popularity of decisions (for Increasing Amounts and Transactions in faraway places) was to mark the transaction as 'fraud', then 'contact customer' and then 'allow'. In these cases, the students were more suspicious of the transactions than the experts.

Interestingly, in the Large Amounts fraud type, both experts and students applied the same strategy of 'contact customer' then 'allow' and 'fraud' as the last resort. In this case, students suggested that there was insufficient information on the dashboard and so they felt that calling the customer would be appropriate.

Across both groups, students and experts, once a strategy was applied to a fraud type, it did not seem to change between dashboards. This meant that the analysis was applied in much the same way, even though the display of information differs.

Certainty associated with flagged patterns relate to the confidence of automated scoring system that the pattern in question is fraudulent. This ranges from 0-100% and one can assume that when the confidence score (or certainty) is close to 0, it is highly unlikely that that the pattern is fraudulent and should be allowed. Conversely, when certainty approaches 100%, it is very likely that fraud has occurred and the credit card needs to be blocked in order to prevent further losses. However, there is a middle range where there is no definite answer and to which human expertise can add value. Analysts can investigate further, by requesting more information or contacting the card holder, and when a decision can be made, either allow further transactions or block the card in question. The design of the User Interfaces was intended to support this behaviour, allowing users to drill down to find further information if required. These evaluation studies suggest that, even when the computer is confident in its decision, when users are asked to review this decision, they will seek to explore as much information as they can prior to making their final recommendation.

5. Quantitative Analysis

While the previous sections presented the results of the user interface evaluation, the goal of this section is to measure the SPEEDD prototype in terms of precision, recall and latency for the fraud management use case. We also present results from the automatic fraud pattern construction task.

Complex Event Recognition

The results regarding complex event recognition presented in the section were collected through the generation of a dataset where the defined patterns were closed, in order to allow the engine to properly detect any fraud.

Precision and recall. Given the above dataset, referred as the sample annotated data, the SPEEDD prototype was able to reach a precision of 1 and recall of 0.985. When comparing the uncertain case to the uncertain (certainty threshold > 0.6), the recall was 1 and precision 0.97. There was also a difference in terms of detection speed, comparing the certain and uncertain case. The former was able to produce alerts in average, 3 minutes earlier than the certain counter part.

Latency. In terms of latency, the complex event process engine was evaluated in two different settings. First, PROTON was used as a standalone tool, while for the second case, PROTON was ran on Storm. Different configurations were also tested for the PROTON on Storm project, in order to achieve the best results.

PROTON stand-alone. Regarding PROTON in standalone mode, the engine presented an average latency of 140 milliseconds, with a maximum latency of 2000 milliseconds. It is important to note that in both applications we have the same type of patterns: COUNT and TREND, therefore there is no difference in the results. These results were collected at a rate of injection of 100 events per second.

PROTON On Storm. Regarding PROTON running on top of Storm, due to the number of processing parameters, we were able to test different configurations for different injection scenarios, in order to achieve the best performance. In addition to the improvements in the stand-alone engine, some improvements were also introduced in the Strom version of the engine.

Table 6.1 summarizes our outcomes. It can be conclude that adding more workers improves latency. Using 2 workers and a CEP parallelization factor of 4, it is already possible to achieve lower average latencies for 500 events per second than what was obtained for 100 events per second for the PROTON standalone version.

Table 6.1 - Performance Results Summary (90% percentile values)

Config	Number of workers	CEP parallelization factor	End-to-end latency (ms) 50 events/sec	End-to-end latency (ms) 500 events/sec
1	1	1	31	189
2	1	2	86	253
3	2	4	25	111
4	4	8	14	44.7
5	4	16	16	87
6	8	16	11	16

Fraud Pattern Construction

In this section we present experimental results from the task machine learning fraud patterns from data. For our experiments we used OLED (Online Learning of Event Definitions) [26], an online system for learning logical rules based on Inductive Logic Programming (ILP) [27]. Due to privacy reasons we were not able to perform a sufficient number of experiments on the real data that FeedZai possesses. We therefore used a synthetic dataset, created by Feedzai. The fraud occurrences in this dataset include instances of the following fraud patterns:

- The “increasing/decreasing amounts” pattern, where fraudulent behaviour is inferred if a number of consecutive transactions occur for a particular card within a small period of time, where the amount of each transaction is respectively larger/smaller than the amount of the previous one.
- The “big-after-small'” pattern, where a withdrawal of a large amount from a particular card follows the withdrawal of a very small amount from the same card, within a small period of time.
- The “flash attack'” pattern, where a large number of transactions with the same card occur within a small time period.
- The “far-away locations” pattern, where two transactions occur with the same card, within a small time period, where that respective acquire banks differ.
- The “card expires” pattern, where a transaction occurs too close (e.g. a day before) to a card's expiration date.

Table 6.1 - Performance comparison: OLED vs a batch ILP learner

Approach	F1-score	Precision	Recall	Time (minutes)
OLED	0.830	0.894	0.776	21
SC	0.892	0.912	0.874	188

In addition to instances of fraudulent transactions, the dataset contains non-fraudulent sequences.These sequences were generated using a predefined “genuine card” pattern. Positive examples in the dataset we used for our experiments with amount to 0.2% of the dataset (the remaining dataset consists of negative examples). This is in accordance to the positive/negative example ratio in Feedzai's real dataset. Moreover, this imbalance of positive and negative examples makes the learning task very challenging. An additional challenge is that fraud patterns often consist of long transaction sequences. This intensifies the task of learning rules for such patterns, since the complexity of a rule increases with its length. The dataset consisted of one million transactions, which amounts to approximately 200 MBs of data.

In our experiments we followed a windowing approach. The training set was consumed in the form of data batches of a pre-defined time-span (windows), where the length of the windows ranged from a few minutes to one day. The goal of our first experiment was to assess the tradeoff between efficiency and quality of the outcome, due to OLED's online nature. To this end, we compared OLED to a classic, offline (batch) ILP learner, which learns one rule at a time in a standard set cover loop [27] that requires several passes over the data. To this end, we implemented such an offline algorithm. The reason for not using one of the existing ILP learners for this task, see [27] for a review of such learners, is that these systems do not support learning from batches of logical atoms, but they instead accept training examples in the form of single logical atoms.

We performed 10-fold cross-validation with both systems (OLED and the batch ILP algorithm). In each run of the cross-validation process, 90% of the data were used for training, while the remaining 10% was retained for testing. We measured average training time, as well as average F1-score, which was calculated by micro-averaging the results of each fold. To ensure that every testing set in each fold has a number of positive examples for evaluation, we split the positives into ten chunks of approximately equal size. In each fold, nine of these chunks were added to the training set (along with 90% of the negative examples), while one was added to the testing set (along with the remaining 10% of the negative examples). Therefore the positive/negative example ratio in the training/testing sets of each fold was similar to the positive/negative ratio in the entire dataset (where positives are 0.2% of the training data). All experiments were conducted on a Linux computer with 8 Intel i7-4770 cores at 3.40GHz and 16 GBs of RAM. Both OLED and the batch ILP algorithm were implemented in the Scala programming language using the Clingo answer set solver[2] as the main reasoning component.

[2] http://potassco.sourceforge.net/

Figure 17 presents the results of the first experiment, where by SSC we denote OLED's set-cover-based rival. SSC achieved better accuracy as compared to OLED. This was expected, since each rule learnt by SSC is highly optimised over the entire dataset. The downside is that SSC's average running time is larger than 3 hours, in contrast to OLED, which learns a collection of fraud patterns of comparable quality in approximately 21 minutes. Note that these times were obtained with a highly parallel implementation of SSC, where the expensive task of repeatedly evaluating candidate rules over the entire training set was split across all available cores. In contrast, OLED used a single core for learning, which processed the entire stream. This is because a highly parallel version of OLED is still under development and simply splitting parts of its functionality that are easily parallelizable (e.g. rule evaluation) did not yield any significant speed-ups in training time.

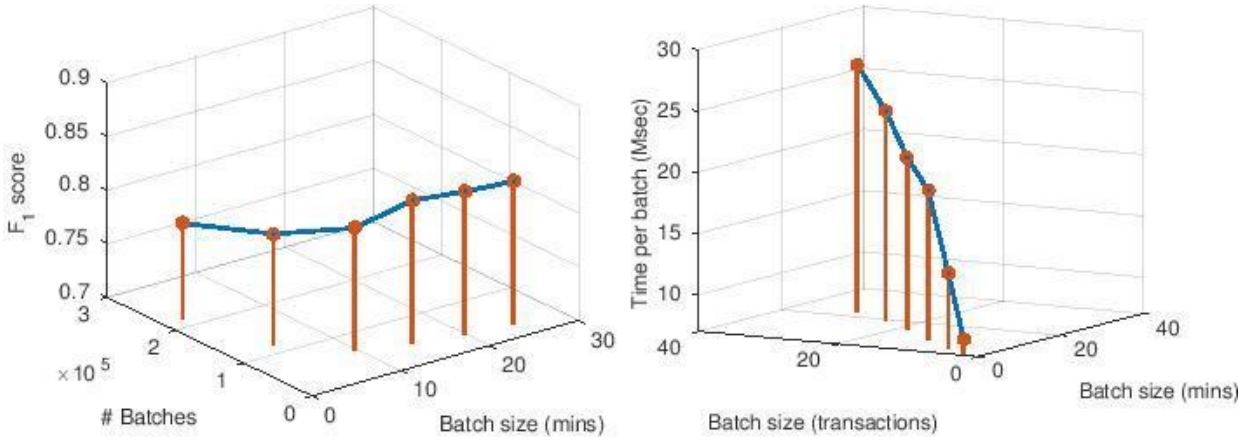

Figure 18. F1 score (left) and average processing time per batch (right) for OLED, for data batches of length 2.5, 5, 10, 15 and 20 minutes.

In our second experiment we studied how the quality of the outcome (in terms of F1-score) and the average processing time per batch are affected in OLED by varying the batch size. To this end, we first conducted experiments with windows of 2.5, 5, 10, 15, 20, 25 minutes. The results are presented in Figure 18. The left-most graph in Figure 18 presents F1-score as a function of batch (window) size in minutes and the total number of batches in the training set, for the particular batch size. Each F1-score value is a (micro-) average obtained from a 5-fold cross-validation process for a particular batch size. The right-most graph in Figure 18 presents the average (over 5-fold cross-validation) processing time per batch as a function of the average number of transactions per batch and batch size in minutes. The experimental setting for each fold of the cross-validation process (positive/negative example ratio in the training/testing sets) was identical to the one described in our previous experiment. Regarding F1-score, results indicate that it grows for window sizes up to 15 minutes, while it remains almost constant for larger window sizes. The reason for that is that windows of size smaller than 15 minutes often contain only part of the transactions that are involved in a particular fraudulent behavior. As a

result, OLED learns incomplete patterns of lower quality when the training examples are presented in windows that are too small. Regarding average processing time per batch, our results indicate that it grows almost linearly with window size.

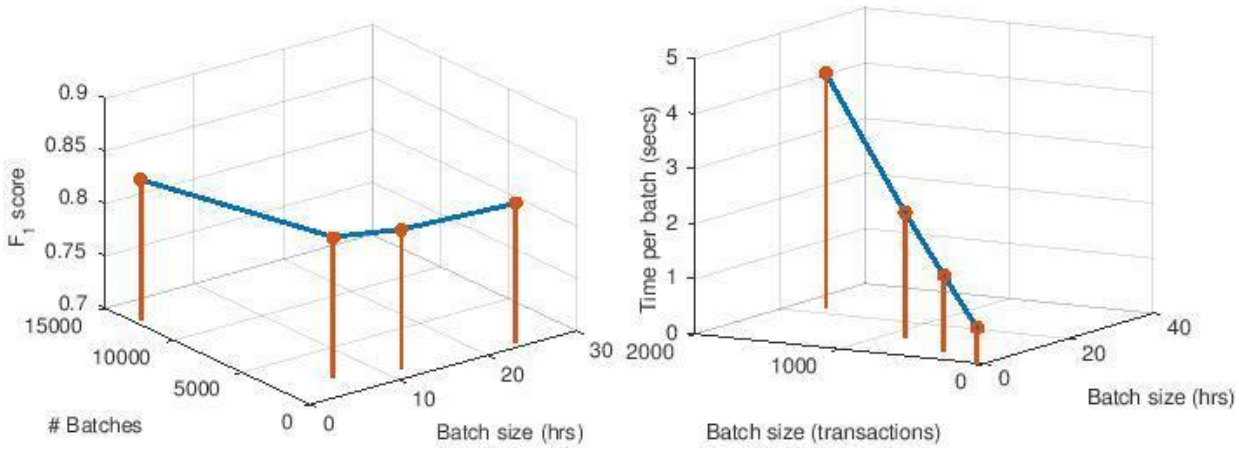

Figure 19. F1 score (left) and average processing time per batch (right) for OLED, for data batches of length 1, 6, 12 and 24 hours respectively.

To stress-test OLED, we also performed similar experiments with larger window sizes of up to one day, in particular 1, 6, 12 and 24 hours. The results are presented in Figure 19. The F1-score remains almost constant for varying window sizes. Regarding average processing time per batch, while it grows significantly, as compared to the previous experiment (Figure 18), it still grows linearly with window size.

6. Discussion

The report has presented two studies in which the SPEEDD dashboards for the credit card fraud use case have been evaluated. In the first study, a cohort of students showed significant performance advantages (in terms of decision time) when using the Detailed (2) dashboard. The experts showed a decision time of around 18s, which was a little slower than the students (14s overall). This could be due to their familiarity with the domain (and the type of information presented to them) and the desire to ensure that the information is used appropriately, or to the use of a 'contact cardholder' as a default response. This latter, if it is the case is interesting in that the fraud analysts' behaviour would imply that removing the human from the loop is not possible at present. Rather, the role of the fraud analytics would be screen transactions to such an extent that the experienced fraud analyst is able to review those cases which have been flagged and then engage in further exploration in order to reach a decision. This is not to deny the importance of automated analysis of the vast majority of transactions. However, when the transactions do not precisely meet the definitions in the algorithms, then there may remain a need for human scrutiny. Consequently, dashboards which can support easy and rapid interpretation of the available information can be beneficial for analysts.

In terms of the baseline decision time, it was proposed that a typical decision by a call handler might take around 2 minutes (assuming 200 cases to investigate in an 8 hour working day, with 1 hour of breaks). Given that the handling of a case would involve blocking or allowing the transaction (and so involve completion of forms for audit purpose) or speaking to the cardholder prior to making the decision (and so involve a telephone call as well as form filling), then one might anticipate the decision on the fraud (based on the information provided) to be performed relatively quickly. Consequently, having a system that supports quick but accurate decision making could be advantageous.

Regarding the quantitative analysis, it was observed that given a rich set of fraud patterns defined in the engine, high values for precision and recall can be obtained. More importantly, is the fact that in the uncertain case, the engine was able able to alert in an average 3 minutes before its counterpart in the certain case, thus giving enough time to an operator to block a credit card earlier than they can today.

7. References

[1] Leonard, K.J. (1993) Detecting credit card fraud using expert systems, *Computers & Industrial Engineering, 25*, 103-106.
[2] Sahin, Y., Bulkan, S. and Duman, E. (2013) A cost-sensitive decision tree approach for fraud detection, *Expert Systems with Applications*, *40*, 5916-5923.
[3] Duman, E. and Ozcelik, M.H. (2011) Detecting credit card fraud by genetic algorithm and scatter search, *Expert Systems with Applications*, *38*, 13057-13063.
[4] Bhattacharyya, S., Jha, S., Tharakunnel, K. and Westland, J.C. (2011) Data mining for credit card fraud: A comparative study, *Decision Support Systems, 50*, 602-613
[5] Hand, D.J. and Blunt, G. (2001) Prospecting for gems in credit card data, *IMA Journal of Management Mathematics, 12*, 173-200
[6] Dilla, W.N. and Rascke, R.L. (2015) Data visualization for fraud detection: practice implications and a call for future research, *International Journal of Accounting Information Systems, 16,* 1-22.
[7] Few, S. (2007) Data Visualization Past, Present and Future, Cognos Innovation Center for Performance Management, [https://www.perceptualedge.com/articles/Whitepapers/Data_Visualization.pdf, accessed 20092016]
[8] 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
[9] Linden, I. (2014) Proposals for the integration of interactive dashboards in business process monitoring to support resources allocation decisions, Journal of Decision Systems, 23, 318-332
[10] Mosier, K.L., Skitka, L.J., Heers, S. and Burdick, M. (1998) Automation Bias: Decision Making and Performance in High-Tech Cockpits, *International Journal of Aviation Psychology,* 8, 47–63
[11] Lee, J.D. and Moray, N. (1994) Trust, self-confidence, and operators' adaptation to automation, *International Journal of Human-Computer Studies,* 40, 153–184
[12] K. Goddard, A. Roudsari, and J. C. Wyatt, "Automation bias: a systematic review of frequency, effect mediators, and mitigators,"
[13] Westin, C., Borst, C. and Hilburn,B. (2016) Strategic conformance: overcoming acceptance issues of decision aiding automation? *IEEE Transactions on Human-Machine Systems, 46,* 41-52.
[14] L. J. SKITKA, K. MOSIER, and M. D. BURDICK, "Accountability and automation bias," Int. J. Hum.-Comput. Stud., vol. 52, no. 4, pp. 701–717, Apr. 2000.
[15] Parasuraman, R. and Manzey, D.H. (2010) Complacency and Bias in Human Use of Automation: An Attentional Integration, *Human Factors*, *52*, 381–410
[16] Parasuraman, R. and Riley, V. (1997) Humans and Automation: Use, Misuse, Disuse, Abuse, *Human Factors, 39*, 230–253
[17] J. E. Bahner, A.-D. Hüper, and D. Manzey, "Misuse of automated decision aids: Complacency, automation bias and the impact of training experience," *Int. J. Hum.-Comput. Stud.*, vol. 66, no. 9, pp. 688–699, Sep. 2008.
[18] C. D. Wickens, B. A. Clegg, A. Z. Vieane, and A. L. Sebok, "Complacency and Automation Bias in the Use of Imperfect Automation," *Hum. Factors J. Hum. Factors Ergon. Soc.*, p. 0018720815581940, Apr. 2015.
[19] Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G. and Beck, H.P. (2003) The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*, 697–718
[20] Tufte, Edward R. (1983) *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press.
[21] Harris, Robert L. (1999) *Information Graphics: A Comprehensive Illustrated Reference*. USA: Oxford University Press
[22] Cohen, J. (1988) *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence

Erlbaum
[23] Hart, S. G., & Staveland, L. E. (1988) Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research, *Advances in psychology*, *52*, 139-183.
[24] Hall, C.C., Ariss, L., Todorov, A. (2007) The illusion of knowledge: when more information reduces accuracy and increases confidence, *Organizational Behaviour and Human Decision Processes, 103,* 277-290.
[25] Oskamp, S. (1965) Overconfidence in case-study judgments, *Journal of Consulting Psychology, 29*, 261–265.
[26] Katzouris,N., Artikis A., and Paliouras G..Online Learning of Event Definitions. *Theory and Practice of Logic Programming* 16.5-6 (2016): 817-33.
[27] Raedt, Luc De. *Logical and Relational Learning*. Berlin: Springer, 2008.