This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

End-User Development for Interactive Data Analytics: Uncertainty, Correlation and User Confidence

Jianlong Zhou, Syed Z. Arshad, Xiuying Wang, Zhidong Li, Dagan Feng, and Fang Chen

Abstract— This paper investigates End-User Development (EUD) for interactive data-analytic interfaces – building upon the ideas of making machine learning transparent. The research is carried out in a business operation environment (water pipe failure prediction in our case) motivated to integrate advanced analytics into decision-making processes of an urban Internet of Things (IoT) concept. We explore effects of revealing uncertainty and correlation on user confidence in a data-driven decision making scenario. It was found that user confidence varied significantly amongst various user groups when different machine learning models were displayed with/without supplementary information. Galvanic Skin Response (GSR) signals were analyzed and shown as reasonable indices for predicting user confidence levels. Supplementary data visualizations (of inherent uncertainty and correlation in data) contributed to explicability principles while GSR indexing added towards correctibility principles. We recommend transparent machine learning as the key to effective EUD for interactive data analytics.

Index Terms— Correlation, decision making, Galvanic Skin Response, machine learning, uncertainty, user confidence

1 INTRODUCTION

THE Internet of Things (IoT) helps people to improve their experiences in their work and life by seamlessly integrating a large number of smart objects with the Internet [1]. Human-Computer interface, which is one of three important foundations of IoT [2], promises to further integrate computing devices into human environment. In these interfaces, the computer anticipates the needs as it learns and grows from the evolving history of interaction with the human. The learning here refers to automated Machine Learning (ML). These interfaces would be better contextually aware of both the user and the environment.

We consider End-User Development (EUD) to be the natural consequence of the IoT condition. Users would like their user interfaces to be flexible enough to accommodate any changing user requirements. Moreover they might also like this evolutionary process to be friendly enough that users can program these changes into their interface whenever needs arise. Burnett and Kulesza [3] argued to "enable ordinary users to customize, control and 'fix' Internet of Things applications that are trying to help them". And the approach they have worked on is known as "explanatory debugging" [4]. Explanatory debugging is an interactive machine learning approach in which the system explains to users how it made each of its predictions (i.e. the explainability principle), and the user then explains any necessary corrections back to the learning system (i.e. the correctibility principle). This effort can be considered as part of making ML transparent. However we focus on the improved decision making by uncertainty communication through physiological signal analysis that helps the system understand the user confidence in real-time.

The role of human in IoT is no longer that of a simple user. Things that differentiate a person participating in IoT from a traditional cyber-physical component are their (a) cognitive abilities (b) unpredictability and (c) motivational factors [5]. In this study we restrict ourselves to explore the cognitive abilities of the human with regards to uncertainty and correlation visualizations and how predictable would be their decision making behavior with respect to ensuing user confidence. This in turn helps us recommend valuable parameters for end-user development design.

Furthermore, much of data analytics in IoT is based on ML techniques that make use of data from various IoT sensors to make IoT appear "smart" in decision making [6]. By attempting to make the ML procedures transparent, we hope to provide users greater insights into the rationale behind the solutions made by IoT systems and thereby improve mutual trust between the user and systems. Once the user realizes how sure/certain a system is of its analysis and recommendations – there would lesser cases of

- *J. Zhou, S.Z. Arshad, Z. Li, and F. Chen are with DATA61, CSIRO, 13 Garden Street, Eveleigh, NSW 2015, Australia, Email: {jianlong.zhou, syed.arshad, zhidong.li, fang.chen}@data61.csiro.au.*
- *X. Wang and D. Feng are with School of Information Technologies, The University of Sydney, NSW 2006, Australia, Email: {xiu.wang, dagan.feng}@ sydney.edu.au.*

*Please note that all acknowledgments should be placed at the end of the paper, before the bibliography (**note that corresponding authorship is not noted in affiliation box, but in acknowledgment section**).*

xxxx-xxxx/0x/$xx.00 © 200x IEEE Published by the IEEE Computer Society

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

overconfidence or underconfidence in system performance, thereby setting upper and lower bounds for end-user development scenarios.

Fig. 1 explains the bigger picture for EUD for interactive data analytics with making ML transparent as the key-contributing factor in this paper. In this scheme, cognitive computing tries to make computers more user-friendly, and anticipates what the user is trying to do (by monitoring the user generated signals and environment) as well as provides an appropriate response. This information contextualizes the potential range of responses, which are therefore more personalized [7]. Such personalization is either deliberately done by the user or gradually learned over time using ML over user interaction data. EUD in this scheme is seen as an extension of personalization. Meanwhile the decision analytics interface is powered by the automated ML techniques, which continuously learn from the raw IoT data. The key here is the bidirectional connection between personalization and automated ML. Personalization helps guide how analytics is to be presented and transformed to the user, whereas ML of interaction data helps customize the personalization component. *Transparency* is when both are able to benefit each other for the achievement of the common system goal of better decision making. Based on this scheme, we give more background on IoT and humans in data analytics-driven decision making in the following subsections.

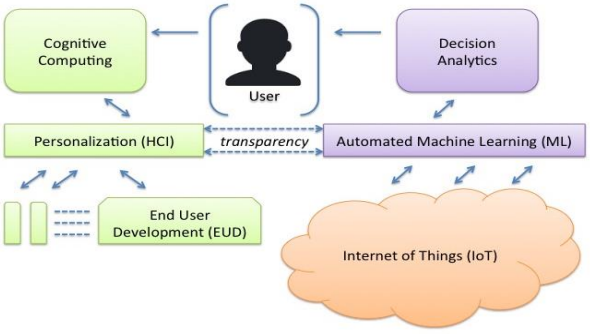

Fig. 1. Bigger picture of EUD for interactive data analytics.

1.1 Humans in the Loop and Decision Making

IoT is inherently about human [3], [8] – providing information to human and getting feedback from human. Therefore, the human factors on the IoT system play significant roles on the success of the IoT. Furthermore, the multimodal interface trend in Human-Computer Interaction (HCI) [9] tries to build interfaces intelligent enough to actively incorporate user's intuitions and load. In case of interactive data analytics, the key HCI research questions would be (see Fig. 2): (a) what aspects of data would users like to see on screen? (b) how could the desired data aspects be best visualized? and (c) how much control can be transferred for the user to adequately manipulate the visualized data? Here we concern ourselves mainly with the first two questions.

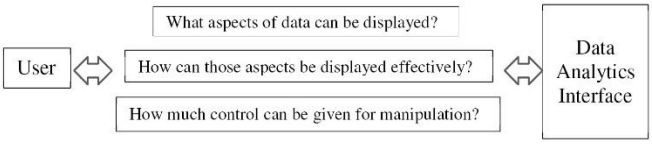

Fig. 2. Interactive data analytics: An HCI perspective.

Decision making is an important research topic in HCI with the fast growing use of intelligent systems. With rapidly increasing data in fields such as infrastructure and society, users are looking to integrate their "Big Data" and advanced analytics into business operations in order to become more analytics-driven in their decision making. As a result, we continuously find ourselves coming across ML-based appealing viewgraphs and other predictions that seem to work (or have worked) surprisingly well in practical scenarios. This popularity of machine learning and predictive analytics has created a growing demand for similar tools in non-computing communities besides ML experts. People with no background in ML, would also like to use these powerful techniques to their benefit such as in IoT.

1.2 Data Analytics and IoT

Recently, the application of the IoT paradigm to an urban context is of particular interest under the Smart City concept [10], which is called urban IoT. Urban IoTs are designed to support the Smart City vision, which aims at supporting added-value services for the administration of the city and for the citizens. For example, the infrastructure management such as water pipe failures management is one of significant aspects in the Smart City. Proper maintenance of water pipes of a city requires the continuous monitoring of the actual conditions of pipes and identification of the pipes that are most subject to failures. By analyzing the collected data with machine learning, the urban IoT may provide predictions on pipe conditions in order to make decisions of management and maintenance of pipes.

1.3 EUD for Interactive Data Analytics

Different from information visualization for data understanding [11], ML-based predictive analytics is like a "black box" for many of non-ML users, to which they simply provide their source data and colorful viewgraphs and/or recommendations are displayed as output [12]. The user is more or less unconfident in the ML model output when making decisions based on the ML model output and thus also unconfident in the ML models themselves. Therefore it is highly critical to know how the information presented in the user interface on data and ML models affect user confidence in order to make effective decisions.

Fig. 3 shows a typical ML-based data analysis pipeline. In this pipeline, users need to consider how certain or uncertain of the prediction results are when making decisions. Furthermore, from the input data perspective, statistical information of data such as correlation between variables can describe how much target values are related to features in input data of the model. The correlation may affect us-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

ers' decision making based on their domain experiences, e.g. users may have experiences that the older the water pipes are, the higher the failure rate is. The user might be risking too much by ignoring uncertainties or correlations, while over-conservative safety certification or having low confidence could possibly be wastage of the incredible potential of ML.

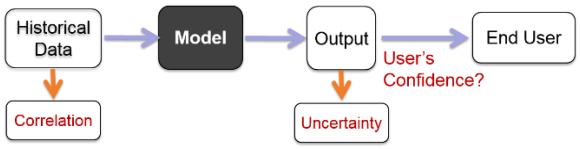

Fig. 3. ML-based data analysis pipeline.

Furthermore, it was found that Galvanic Skin Response (GSR), which corresponds to the electrical conductance of the skin, as a low-cost and robust physiological signal, reflects the process of decision making, in particular, of the emotional sanctioning of an active go-ahead [13]. GSR is often used as an indicator of affective processes and emotional arousal. Many scientific findings indicate that emotions play an essential role in decision making, and various cognitive tasks. These motivate us to investigate how physiological signals such as GSR are used to communicate confidence in data analytics-driven decision making under varying uncertainty and correlation conditions.

However, little work is done on the EUD for interactive data analytics in the IoT by investigating user confidence in data analytics-driven decision making under conditions such as the revealing of uncertainty and correlation. Such investigation will benefit the effective EUD for interactive data analytics by setting up the communication between machine learning and users as shown in Fig. 1. This paper aims to investigate relationships between uncertainty/correlation and user confidence in data analytics-driven decision making in the urban IoT in order to design effective user interface for interactive data analytics systems. We also investigate GSR features that can be used to communicate user confidence in decision making. Different from other HCI communication approaches such as semiotic engineering for communication between designers and users at interaction time [14], our physiological signal-based user confidence communication allows to adapt the presentation of decision conditions automatically for effective decision making. Such communication helps the design of intelligent user interface in IoT where user confidence is automatically identified and perceived explicitly. Based on the investigation, we demonstrate that the transparent machine learning is the key of the EUD for interactive data analytics in IoT. A user study was performed to investigate the impact of revealing uncertainty and correlation information on user confidence.

2 Related Work

2.1 User Uncertainty and Decision Making

Making decisions is one of the most complex cognitive processes. For example, Morgado et al. [15] reviewed the impact of stress in decision making in the context of uncertainty and found that this cognitive process involves several sequential steps including analysis of internal and external states, valuation of options available and action selection.

"Uncertainty" is defined in many ways. For a user, it can be a psychological state in which the decision maker lacks knowledge about what outcome will follow from which choice. This aspect of uncertainty is popularly known as "risk". Risk refers to situations with a known distribution of possible outcomes (probabilities) [16]. "Ambiguity" is the other kind of uncertainty, where outcomes have unknown probabilities and research in neurosciences [17] indicates that decision making under ambiguity does not represent a special, more complex case of risky decision making.

It was thought that people prefer to bet on events they know more about, even when their beliefs are held constant, i.e. they are averse to ambiguity [18]. However, this was shown to be otherwise by [19] in their study responding to degrees of uncertainty. Their experiments and corresponding neurological observations showed that many people are more willing to bet on risky outcomes than ambiguous ones. These findings motivate us to account for both risk (i.e. uncertainty due to known probabilities) and ambiguity (i.e. uncertainty due to unknown probabilities) while investigating variations in user confidence due to uncertainty.

2.2 Presenting Model Uncertainty

Probability remains the language of uncertainty. Several investigations have been carried out to understand better ways of presenting uncertainty inherent in data. One earlier effort includes that of Ibrekk and Morgan [20] who investigated graphical communication of uncertain quantities to well-educated semi and non-technical people. It was suggested that communicating uncertainty should not focus just on the problems of communicating to "semi technical and lay people". We pay heed to this advice and the subject groups we consider in our experiment involve both experts (ML & non-ML) and general staff.

Furthermore, cognitive load is known to affect people's use of graphical displays. Allen et al. [21] studied the potential of graphical displays to communicate uncertainty when end users were under cognitive load. The research suggested that interpreting basic characteristics (like "point reading") of uncertainty data is unharmed under conditions of limited cognitive resources, whereas more deliberative processing, like synthesizing information, is negatively affected. "Synthesizing" here corresponds to cognitive capacity demanding Type 2 processing as explained by Stanovich and Toplak [22]. Type 2 processing is further discussed later. In our current design we maintain cognitive load to be same over all experimental conditions and keep sessions times minimal. We do this because [23] has

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

shown that individual cognitive load can vary over longer periods of activity which is then reflected in changing behavioral patterns.

2.3 Correlation and Decision Making

Good decision-making often requires people to perceive and handle a myriad of statistical correlations [24]. However, Eyster and G. Weizsacker [24] found that people have limited attention and often neglect correlations in financial decision making. Ye [25] used the weighted correlation coefficients to rank the alternatives and get the best alternative in multi-attribute decision making. Liao et al. [26] used correlation coefficients of hesitant fuzzy linguistic term set in the process of qualitative decision making in traditional Chinese medical diagnosis. However, little work is done on how correlation affects user confidence, especially decision making based on ML models learned from historical data.

3 Experiment

3.1 Case Study

This research used water pipe failure prediction as a case study [27]. Water supply networks constitute one of the most crucial and valuable urban assets. Identifying an accurate predictive measure for imminent failure of water pipes would allow utility companies to prioritize preventive repairs that would cost significantly less than full-scale failures. Thus, utility companies use outcomes from failure prediction models, to make renewal plans based on risk levels of pipes and also reasonable budget plans for pipe maintenance. However, different models based on alternative features may be available resulting in different possible budget plans. This experiment is set up to determine what criteria of choice are in favor of a model, and what parameters influence the user confidence during the decision process.

3.2 Experiment Data

Water pipe failure prediction uses historical pipe failure data to predict future failure rate [27]. The historical data contain failure records of water pipes, and various attributes of water pipes, such as laid year, length, diameter size, surrounding soil type, etc. Actual historical data was sampled and customized for the simulation of this experiment.

In this study, predictive models are simulated and they are based on different pipe features (e.g. size) with the reference of Hierarchical Beta Process (HBP) used in water pipe failure prediction [27]. The model performance curve was presented to let the participants evaluate different models. The model performance is the functional relationship between proportion of the network inspected and the proportion of pipe failures detected. Fig. 4 shows the performances of two models. For example, in Fig. 4 (a), the model based on the feature "Size" has better performance than the one based on the feature "Laid Year", because the formal one detects more failures than the latter for a given pipe length.

In machine learning, uncertainty can be traced to many sources ranging from input values to nature of model representation to final output decision values. Here we concern ourselves mostly with uncertainty associated with output decision values. Specifically, model uncertainty here refers to an interval within which the true value of a measured quantity would lie. For example, in Fig. 4 (b), in order to detect 50% of the failure rate, the uncertainty interval of the inspected length is $[5\%, 55\%]$ for the model based on the feature "Size" (it is also called model "Size" for short and similar to other models), and is $[35\%, 45\%]$ for model "Laid Year": the model "Laid Year" is said to have less uncertainty in prediction than the model "Size" because the former has smaller uncertainty interval than the later. Model output uncertainty usually spans as a band in the model performance diagram as shown in Fig. 4 (b) and Fig. 4 (c). By considering model output uncertainty, the relationship between two models may have two cases as shown in Fig. 4: 1) models with overlapping uncertainty, and we call overlapping models (see Fig. 4 (b)), and 2) models with non-overlapping uncertainty, and we call non-overlapping models (see Fig. 4 (c)). In Fig. 4 (b), the model with lower uncertainty overlaps completely the model with higher uncertainty, whereas in Fig. 4 (c), the two bands are disjoint.

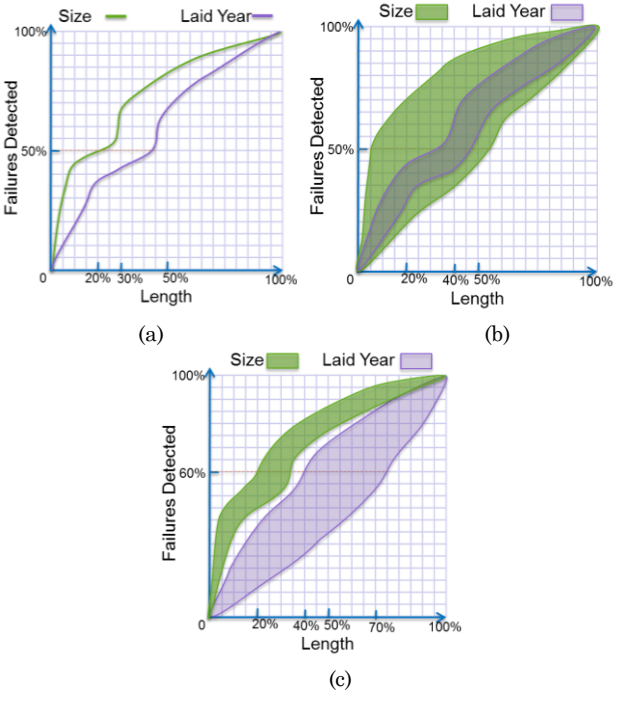

Fig. 4. Performance of predictive models: (a) without uncertainty, (b) with overlapping uncertainty, (c) with non-overlapping uncertainty.

Furthermore, as represented in Fig. 3, correlation is not associated with a model, but with input data. Correlation in this experiment refers to the correlation between one pipe feature (e.g. pipe size) and the pipe failure rate in historical records. It describes how much the target value "failure rate" is related to a given feature in historical records. The correlation is often described by correlation coef-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

ficient. It illustrates a quantitative measure of correlation and dependence. The correlation in this experiment is displayed as 2D bar charts with the horizontal axis as features and the vertical axis as the correlation coefficients. For example, in Fig. 5, the feature "Laid Year" (Year) and "Size" have a correlation coefficient of 0.75 and 0.45 respectively with the failure rate, which means that the feature "Laid Year" is more related to the failure rate than the feature "Size".

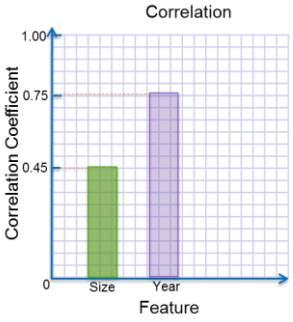

Fig. 5. Correlation between features and target values.

3.3 Task Design

In this subsection, tasks are designed to investigate how uncertainty and correlation affect user confidence in decision making. In each task, different diagrams for two ML models with other decision making information are presented to participants for decisions. Based on diagrams presented to participants, the tasks are divided into four categories: control tasks, uncertainty-based tasks, correlation-based tasks, and combinational tasks.

3.3.1 *Control Tasks*

A Control Task (CT) is used as the basis to test the effects of uncertainty and correlation on user confidence. In the CT task, only the model performance diagram (Fig. 4 (a)) without uncertainty/correlation information is presented.

3.3.2 *Uncertainty-Based Tasks*

Uncertainty-based tasks are used to evaluate the effects of different uncertainty patterns on user confidence. Based on two cases of relationships between two models under uncertainty as shown in Fig. 4 (b) and Fig. 4 (c), two uncertainty tasks are designed: (1) Overlapping Uncertainty Task (OLUT), where only the model performance diagram with overlapping uncertainty is presented to participants (see Fig. 4 (b)); (2) Non-Overlapping Uncertainty Task (Non-OLUT), where only the model performance diagram with non-overlapping uncertainty is presented to participants (see Fig. 4 (b)).

3.3.3 *Correlation-Based Tasks*

Correlation-based tasks are used to evaluate how different correlation patterns affect user confidence in decision making. Considering model performance and correlation together, we divided relations between them into two categories:

- Correlation and performance of model output share the same trend (Fig. 6 (a)). That is, the correlation between a feature and the pipe failure rate is high and the associated model performance is also high, or the contrary.
- Correlation and performance of model output do not share the same trend (Fig. 6 (b)). That is, the correlation is high, but the associated model performance is low, or the contrary.

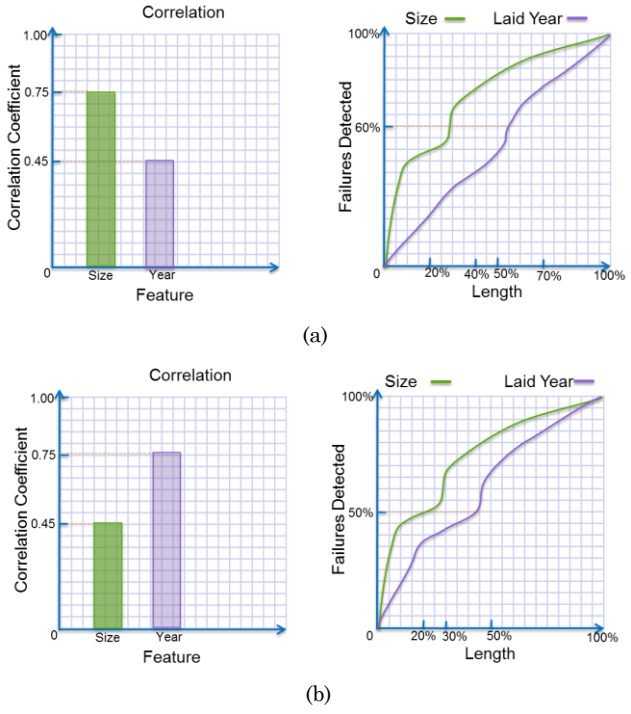

Fig. 6. Correlation-based tasks: (a) Same Trend Correlation Task, (b) Non-Same Trend Correlation Task.

According to these categories, two correlation-based tasks are designed in this study: (1) Same Trend Correlation Task (STCT) (see Fig. 6 (a)); (2) Non-Same Trend Correlation Task (Non-STCT) (see Fig. 6 (b)).

3.3.4 *Combinational Tasks*

Combinational tasks are used to investigate user confidence in decision making when both uncertainty and correlation are presented to participants during decision making. According to two uncertainty cases and two correlation cases as mentioned, four combinational tasks are designed in this study:

- Non-Same Trend Correlation and Overlapping Uncertainty Task (Non-STC+OLU), where model performance with overlapping uncertainty and correlation do not share the same trend;
- Same Trend Correlation and Overlapping Uncertainty Task (STC+OLU), where model performance with overlapping uncertainty and correlation share the same trend;
- Non-Same Trend Correlation and Non-Overlapping Uncertainty Task (Non-STC+Non-OLU), where model performance with non-overlapping uncertainty and correlation do not share the same trend;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

- Same Trend Correlation and Non-Overlapping Uncertainty Task (STC+Non-OLU), where model performance with non-overlapping uncertainty and correlation share the same trend.

3.3.5 Task Procedures

According to the water pipe failure prediction framework, the decision tasks we investigated on are: each user was told that he/she would be a manager of a water company. The water company plans to repair XX% (the exact number was provided during the task) pipe failures in the next financial year. He/she was asked to make a budget plan, i.e. a budget in network length, using water pipe failure prediction models learned from the historical water pipe failure records. Two ML models were provided for each budget task. Participants were required to make decisions by selecting one of presented ML models and then making budget plan based on the selected ML model. The budget plan needs to meet following requirements:

- Check as short length of pipes as possible (low cost);
- The budget uncertainty interval should be as small as possible (high accuracy).

Participants' budget plan was required to report the following information:

- The length of pipes to be detected;
- The prediction model used for the decision.

All tasks were conducted with two rounds. The first round used the feature pair of "Size – Laid Year" and the second round used the feature pair of "Material – Pressure" for ML models. Except feature name differences, two rounds used same model performance diagrams to avoid any bias. In summary, there were 18 tasks conducted ([1 Control Task + 2 Uncertainty-based Tasks + 2 Correlations-based Tasks + 4 Combinational Tasks] × 2 Rounds = 18 Tasks).

At the beginning of each decision making task, a blank screen (with an "X" displayed at the centre of the screen) was displayed for 6 seconds in order to allow the participant have a rest and "reset" his/her cognitive load state [28]. Then the participants started a task under various conditions. Participants were told that they were competing against other people to reach the best budget plan in a given time period (1.5 minutes/task) in order to push them to make their efforts for tasks. The task orders were randomized during the experiment.

3.4 Participants and Apparatus

26 participants were recruited from three groups with different background, with the range of ages from twenties to forties and an average age of 30 years: 1) 9 researchers who were doing ML or data mining research (ML researchers), 2) 8 researchers who were not doing ML or data mining research (non-ML researchers), and 3) 9 administrative staff. Of all participants, 9 were females. Educational qualifications were largely postgraduate (13 PhDs, 6 Masters, 4 Bachelors, 3 other).

GSR devices from ProComp Infiniti of Thought Technology Ltd were used to collect skin conductance responses of subjects. GSR sensors were attached to subjects' left hand fingers. All participants were right-handed. Different tasks were presented on a 21-inch Dell monitor with a screen resolution of 1024 by 768 pixels.

3.5 Data Collection

After each decision making task, participants were asked to rate the confidence level of the budget plan they made and the difficulty level of the task using a 9-point Likert scale (1: least difficult/confident, and 9: most difficult/confident). Participants were asked to rate how helpful the presentation of uncertainty/correlation is for decision making. At the end of each round, participants were also asked to rate the usefulness of uncertainty and correlation on helping them more confident in decision making. Participants were also asked to rank tasks according to difficulty levels of tasks. Besides subjective ratings, skin conductance responses of subjects were collected with GSR sensors during the task time.

4 Hypotheses

The following hypotheses are posed in our study:

- For uncertainty effects:
  - Understanding of uncertainty of model output would help users more confident in decision making (H1);
  - Uncertainty patterns would affect user confidence and users would be more confident in decision making under non-overlapping uncertainty than under overlapping uncertainty (H2);
- For correlation effects:
  - Revealing of correlations between features and target values would help users more confident (H3);
  - When correlation and model performance share the same trend, users would be more confident (H4);
- For combinational effects:
  - When correlation and performance of model output shared the same trend, non-overlapping uncertainty would make users more confident (H5);
  - When correlation and performance of model output do not share the same trend, participants would be more confident in models with overlapping uncertainty than with non-overlapping uncertainty (H6);
- For physiological responses:
  - Confidence variations because of the revealing of uncertainty and correlation would result in differences of physiological measurements (H7).

5 Analysis of Subjective Ratings

We performed Friedman tests with post-hoc analysis using Wilcoxon signed-rank tests to analyze the mean differences in participant responses for each category of tasks.

5.1 Analyses of Uncertainty-Based Tasks

Fig. 7 shows average subjective ratings of participants' confidence in decision making in uncertainty-based tasks

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

and the control task. A Friedman test showed that there was a statistically significant difference among the three tasks in confidence levels, $\chi^2(2) = 13.481$, p<.001. The post-hoc Wilcoxon tests with a Bonferroni correction applied resulting in a new significance level set at p<.017 (0.05/3 = 0.017 because we have three conditions/tasks) was applied to find pair-wise differences between tasks.

It was found that users were significantly more confident in Non-OLUT task than in OLUT task (Z=79.0, p<.001). The result suggests that when uncertainty was presented to users, non-overlapping uncertainty made users more confident in decision making than overlapping uncertainty as we expected (H2). However, we did not find significant differences between Control task and OLUT task or between Control task and Non-OLUT task as we expected (H1).

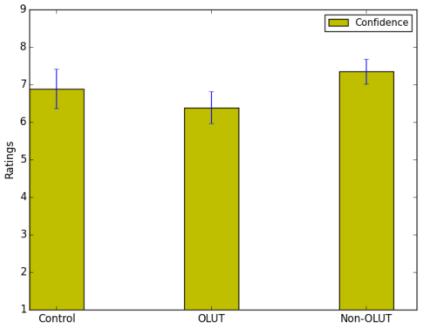

Fig. 7. Average subjective ratings of participants' confidence in decision making in uncertainty-based tasks

5.2 Analyses of Correlation-Based Tasks

A Friedman test found a statistically significant difference among three tasks in confidence levels, $\chi^2(2) = 22.086$, p<.001. The post-hoc Wilcoxon tests (a significance level at p<.017) found a significant difference between Control Task and Same Trend Correlation Task (Z=167.5, p=.008). The result suggests that revealing of correlations between features and target values helped users more confident in decision making as we expected (H3). It was also found that participants were significantly more confident in Same Trend Correlation Task than in Non-Same Trend Correlation Task (Z=105.0, p<.001). The result suggests that when correlation and performance of model output shared the same trend, users were more confident in decision making as we expected (H4).

5.3 Analyses of Combinational Tasks

A Friedman test found a statistically significant difference among four tasks in confidence levels, $\chi^2(3) = 55.886$, p<.001. The post-hoc Wilcoxon tests a new significance level set at p<.0125 (0.05/4=0.0125, because we have four tasks) was then applied to find pair-wise differences between tasks. The post-hoc tests found that participants were significantly more confident in STC+Non-OLU Task than in STC+OLU Task (Z=89.0, p<.001). The result suggests that when correlation and performance of model output shared the same trend, non-overlapping uncertainty made users more confident in decision making as we expected (H5). It was also found that participants were significantly more confident in STC+Non-OLU than in Non-STC+Non-OLU (Z=10.0, p<.001). The result suggests that under the condition of non-overlapping uncertainty, the same trend between correlation and performance of model output made users more confident in decision making as we expected (H5). Participants were also statistically significantly more confident in STC+Non-OLU Task than in Non-STC+OLU Task (Z=103.5, p=.002). It suggests that both the same trend scenario of correlation and the non-overlapping uncertainty benefited user confidence in decision making. The post-hoc tests also found that participants were significantly more confident in Non-STC+OLU Task than in Non-STC+Non-OLU Task (Z=149.5, p<.001). The result suggests that when correlation and performance of model output did not share the same trend, participants were more confident in models with overlapping ("ambiguity") uncertainty than models with non-overlapping ("risk") uncertainty as hypothesized (H6). This is maybe because of the popular assumption of human's risk-aversion in decision making [29].

6 Analysis of GSR Responses

In this section, GSR responses from subjects are analysed. Fig. 8 shows an example of GSR signals of a participant in one task session. Various features are firstly extracted from GSR signals. GSR features are then used to classify confidence levels in order to show the potential of using GSR in indexing user confidence in decision making in the next section. GSR responses during both task time and the period of displaying "X" are used to analyse user confidence. The GSR data analysis is divided into following steps: 1) data calibration, 2) signal smoothing, 3) extrema detection, 4) feature encoding, and 5) feature significance test.

6.1 GSR Features

This subsection shows the steps to extract and encode GSR features in this study. The 6-second GSR values before the task start time during the displaying of "X" are used to calibrate GSR during the task time in order to compensate the differences between tasks of a subject. A Hann window function [30] is convoluted to GSR signals to remove noises. The smoothed signal is also normalized using Z-Normalization to omit subjective differences between various signals before the feature extraction.

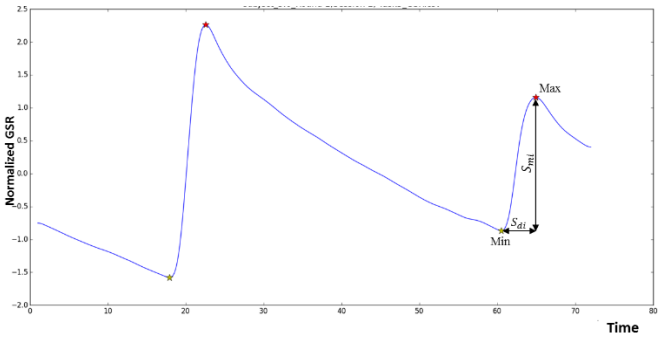

Fig. 8. Extremas and extrema features of GSR.

Both statistical and extrema-based features [31] are extracted and analysed. These features include (see Fig. 8):

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

- Mean of GSR (summation of GSR values over task time divided by task time) μ_G;
- Variance of GSR σ_G;
- Task time length T_t;
- Number of responses S_f, which is the number of peaks in a GSR signal;
- Sum of duration $S_d = \sum S_{di}$;
- Sum of magnitude $S_m = \sum S_{mi}$;
- Sum of estimated area $S_a = \sum S_{ai}$.

S_f, S_d, S_m, and S_a are features of the GSR orienting response [31]. The definition of magnitude S_{mi} and duration S_{di} are defined as shown in Fig. 8. The area of response is estimated by $S_{ai} = S_{mi}S_{di}/2$.

6.2 GSR Feature Significance Test

In this subsection, one-way ANOVA tests with post-hoc analysis using t-tests were performed to evaluate confidence discrimination of features among different tasks.

6.2.1 Uncertainty-Based Tasks

An ANOVA test found that features of T_t ($F(2,39)=4.697$, $p=.013$), S_d ($F(2,39)=3.817$, $p=.029$), S_m ($F(2,39)=3.539$, $p=.036$), and S_a ($F(2,39)=4.52$, $p=.016$) showed statistically significant differences among three tasks (two uncertainty-based tasks plus control task). Post-hoc analysis with t-tests were then conducted with a Bonferroni correction (significance level set at $p<.017$ as discussed in the previous section) for all pairwise differences of significant features. The post-hoc tests showed that OLUT task took significantly longer T_t than Non-OLUT task ($t=2.592$, $p=.014$). There were no other significant differences found between tasks. Furthermore, we used a readjusted significance alpha level of 0.025 (0.05/2 by considering actual two conditions of with/without uncertainty revealing) to see if we can find any other pairwise differences that we expected. Using this new alpha level, the results showed that OLUT task had significantly higher S_d than both Control task ($t=2.353$, $p=.024$) and Non-OLUT task ($t=2.396$, $p=.022$).

The results suggest that overlapping uncertainty made features such as T_t and S_d values increased significantly. Therefore, less confidence level tasks made task time length of T_t and GSR feature S_d values significantly higher.

6.2.2 Correlation-Based Tasks

An ANOVA test showed that GSR features of S_d ($F(2,39)=3.477$, $p=.038$), and S_a ($F(2,39)=4.13$, $p=.021$) showed statistically significant differences among three tasks (two correlation-based tasks plus control task). The post-hoc t-tests with a Bonferroni correction (significance level set at $p<.017$ as discussed in the previous section) showed that Non-STC Task had significantly higher S_d ($t=2.711$, $p=.010$) and S_a ($t=2.889$, $p=.006$) than Control Task. The post-hoc tests with a readjusted significance alpha level of 0.025 further found that Non-STC Task had significantly higher S_m ($t=2.4$, $p=.021$) than Control Task. It was also found that STC task had significantly higher S_a ($t=2.352$, $p=.025$) than Control Task.

These results confirmed our findings in uncertainty-based tasks that lower user confidence levels were correlated to high GSR values of S_d, besides S_m and S_a.

6.2.3 Combinational Tasks

An ANOVA test did not find any significant differences among combinational tasks for all GSR features. However, the pairwise t-tests with a Bonferroni correction (significance level set at $p<.05/2=.025$ by considering that we have two conditions of uncertainty and correlation to be investigated in tasks) showed that Non-STC+Non-OLU task had significantly higher S_m ($t=2.328$, $p=.024$) than Non-STC+OLU task. It was also found that Non-STC+Non-OLU task had significantly higher S_a than both Non-STC+OLU task ($t=2.619$, $p=.012$) and STC+Non-OLU task ($t=2.431$, $p=.021$).

In a word, all results in this section confirmed that tasks with lower user confidence levels were correlated to high values of S_d, S_m, and/or S_a, as well as longer task time length T_t. All these confirm our hypothesis H7.

7 User Confidence Classification Based On GSR Features

This section examines GSR for indexing confidence levels quantitatively. Such indexing can be used in interactive data analytics-driven decision making applications in order to let users perceive their confidence levels in decision making in real-time automatically. Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF), and C4.5 classifiers were applied for classification of confidence levels in decision making based on GSR features. These classifiers are widely used for feature classification in machine learning. SVM can be advantageous for capturing complex relations in data without manual intervention. The Naïve Bayes classifier applies Bayes theorem, which considers each feature to have contributed independently. While it can be trained very efficiently, it nonetheless contains oversimplified assumptions like independence of features. Random Forest combines multiple decision trees during training of data. It predicts the class by combining decisions from individual trees. C4.5 is an algorithm to build decision trees for feature classifications. The leave-one-out method was used in the cross validation. The four classifiers are compared in order to identify the best classifiers in indexing confidence levels.

7.1 Uncertainty-Based Tasks

In uncertainty-based tasks, all seven identified features including four identified significant features (S_d, S_a, S_m, T_t) were used to examine two-class as well as three-class classifications of confidence levels. According to statistical results of subjective ratings as shown in Fig. 7, Non-OLUT Task and Control Task were considered as high confidence level tasks while OLUT Task was used as low confidence level task in two-class classifications. In three-class classifications, Non-OLUT Task was considered as high confidence level task, Control Task was considered as middle confidence level task, and OLUT Task was considered as low confidence level task.

The classification accuracies are shown in Table 1. The

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

results show that both Random Forest and C4.5 with S_d (Accuracy: 72.2%) outperform any other classifiers for two-class classification. Both Random Forest with S_d (Accuracy: 59.3%) and Naïve Bayes with T_t (Accuracy: 59.3%) outperform any other combinations for three-class classification. The results suggest that GSR especially features of S_d and T_t can be used to indicate user confidence levels in decision making effectively.

TABLE 1
User confidence classification accuracies with GSR features in uncertainty-based tasks (All: all GSR features, Sig.: all significant features).

	2-Class						3-Class					
	S_d	S_a	S_m	T_t	*Sig.*	*All*	S_d	S_a	S_m	T_t	*Sig.*	*All*
SVM	0.574	0.611	0.519	0.537	0.500	0.630	0.426	0.370	0.407	0.389	0.407	0.407
RF	**0.722**	0.556	0.519	0.648	0.667	0.611	**0.593**	0.389	0.333	0.519	0.574	0.463
NB	0.704	0.685	0.593	0.593	0.667	0.611	0.519	0.352	0.296	**0.593**	0.482	0.407
C4.5	**0.722**	0.204	0.426	0.593	0.704	0.593	0.482	0.204	0.482	0.574	0.574	0.444

7.2 Correlation-Based Tasks

Similar to the previous subsection, all seven identified features including two identified significant features (S_d, S_a) were used to examine two-class as well as three-class classifications of confidence levels in correlation-based tasks. In two-class classifications, STCT Task and Control Task were considered as high confidence level tasks while Non-STCT Task as low confidence level task in two-class classifications based on subjective ratings. In three-class classifications, STCT Task was considered as high confidence level task, Control Task was considered as middle confidence level task, and Non-STCT Task was considered as low confidence level task based on subjective ratings as discussed before.

The classification accuracies are shown in Table 2. The results show that C4.5 with both S_d and all significant GSR features (Accuracy: 72.2%) outperforms other classifiers in two-class classifications. In three-class classifications, C4.5 with all significant GSR features (Accuracy: 48.3%) outperforms other classifiers. The results confirm that GSR features especially S_d can be used to index user confidence levels in decision making effectively as found in the uncertainty-based tasks.

TABLE 2
User confidence classification accuracies with GSR features in correlation-based tasks (All: all GSR features, Sig.: all significant features).

	2-Class				3-Class			
	S_d	S_a	*Sig.*	*All*	S_d	S_a	*Sig.*	*All*
SVM	0.552	0.379	0.589	0.466	0.362	0.400	0.414	0.448
RF	0.621	0.621	0.552	0.569	0.310	0.448	0.345	0.448
NB	0.638	0.621	0.638	0.552	0.448	0.400	0.431	0.345
C4.5	**0.707**	0.448	**0.707**	0.672	0.417	0.345	**0.483**	0.414

7.3 Combinational Tasks

All seven identified features as well as significant extreme features identified in uncertainty and correlation based tasks were also used to examine two-class as well as three-class classifications of confidence levels in combinational tasks. In two-class classifications, STC+Non-OLU Task and Non-STC+OLU Task were considered as high confidence level tasks, while STC+OLU Task and Non-STC+Non-OLU Task as low confidence level task based on subjective ratings. In three-class classifications, STC+Non-OLU Task was considered as high confidence level task, Non-STC+OLU Task was considered as middle confidence level task, and STC+OLU Task and Non-STC+Non-OLU Task were considered as low confidence level task based on subjective ratings.

TABLE 3
User confidence classification accuracies with GSR features in combinational tasks (All: all GSR features, Ext.: extreme features of S_d, S_a, S_m).

	2-Class					3-Class				
	S_d	S_a	S_m	*Ext.*	*All*	S_d	S_a	S_m	*Ext.*	*All*
SVM	**0.640**	0.302	0.545	0.488	0.500	**0.581**	0.419	0.523	0.523	0.477
RF	0.570	0.570	0.570	0.535	0.454	0.465	0.419	0.442	0.407	0.395
NB	0.605	0.628	0.545	0.593	0.593	0.547	0.488	0.535	0.547	0.454
C4.5	0.570	0.628	0.535	0.500	0.465	0.535	0.326	0.535	0.535	0.349

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

The classification accuracies are shown in Table 3. The results show that SVM with the GSR feature S_d outperforms other classifiers in both two-class (Accuracy: 64.0%) and three-class (Accuracy: 58.1%) classifications. The results also confirm that GSR features especially S_d can be used in indexing user confidence effectively.

By comparison of user confidence classification performance in three task categories, we can see that GSR shows the best performance in uncertainty-based tasks in indexing user confidence. This is maybe because that uncertainty affected user confidence in decision making more than correlation or their combinations.

8 Discussions

This section first looks at the possible interpretation of overlapping and non-overlapping uncertainty presentations and then analyze the behavior and responses of related subject groups accordingly. The possible interpretation of correlation on user confidence is also discussed. This is followed by a discussion regarding some experimental limitations regarding user confidence theoretical framework. Finally, we look implications of this research for End-User Development.

8.1 Overlapping & Non-overlapping Uncertainty Presentations

We are already familiar with a useful distinction that uncertainty with known probabilities is known as risk while uncertainty with unknown probability is referred to as ambiguity. Referring to Fig. 4 (c), we see that uncertainty can be clearly got for each model (at all points) since it is non-overlapping. This results in the model selection very easy. Simply pick a model that detects more failures for least pipe length checked. After model selection, budget estimate is simply a matter of point reading wherever steepest curve tops (the optimum point). Uncertainty in user budget estimate can be matched to model uncertainty depicted by thickness of the line at that point. This type of uncertainty presentation is very close to risk. From a cognitive perspective, making use of this visual presentation should involve more of Type 1 processes that are less demanding of cognitive capacity, holistic, automatic and relatively fast [22].

On the other hand, overlapping uncertainty presentation depicts thick overlapping model lines (Fig. 4(b)). It is no longer clear which model would be the better choice. Much of the decision now depends on the preferences of the user. In case of complete overlap, the model with greater potential gains also is associated with greater potential losses, whereas the model in the middle is more modest in both regards. Trying to make use of this visual, the decision maker comes across ambiguous patches where probability is not exactly known for given particular points. Budget estimation is no longer a point reading task. Thus the user must make some mental valuations of the ambiguous visual and then decide accordingly. Again from a cognitive perspective, making use of this visual presentation would involve more of Type 2 processes that are more demanding of cognitive capacity, analytic, controlled and relatively slow [22].

8.2 Accepting Uncertainty and its Impact

In response to a separate questionnaire in the experiment, all subjects had unanimously agreed to the usefulness of presenting uncertainty as supplementary material. They also agreed to the helpful role of uncertainty whenever it was presented. However, when uncertainty was actually presented, the user confidence generally decreased and the decision making time significantly increased for cases of overlapping uncertainty. Only for non-overlapping uncertainty presentation did time reduce for some subjects as compared to the case of no uncertainty presentation. To understand this peculiar trend we split up the data to investigate tendencies in each subject group. Group-wise data are presented in Fig. 9 to Fig. 12.

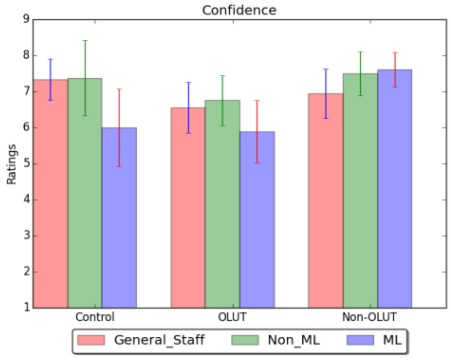

Fig. 9. Average subjective ratings of participants' confidence in decision making by subject groups.

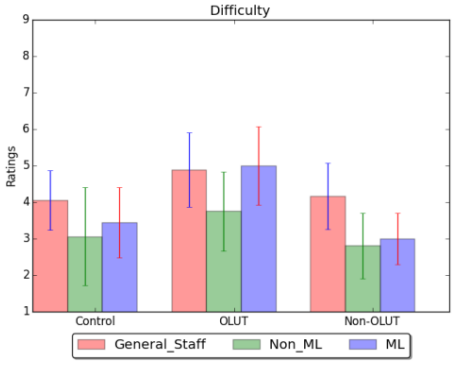

Fig. 10. Average subjective ratings of difficulty levels of decision making tasks by subject groups.

8.2.1 General Staff

Clearly the general staff seemed to be most confident when no uncertainty was presented, then became slightly less confident with non-overlapping uncertainty and finally least confident with overlapping uncertainty scenario (see Fig. 9). This is validated by task difficulty ratings that show general staff found the overlapping uncertainty task to be most difficult (see Fig. 10). It also shows up in user decision making time where general staff seem to take the longest time for overlapping uncertainty tasks (see Fig. 11). Finally this results in large varied budget estimations (see Fig. 12), which can be evidence of general staff not being sure or confident. A possible explanation for this trend could be that although general staff may think uncertainty can be useful in predictive decision making, yet they were not well versed in its usage when

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

uncertainty was actually presented.

8.2.2 Machine Learning Experts/Researchers

Machine Learning experts are the only subject group that seem most confident with non-overlapping uncertainty (see Fig. 9). This is actually the right attitude as non-overlapping uncertainty scenario clearly communicates the inherent uncertainty – making the choice of model very easy and budget estimation just a matter of point reading. This is also matched by the lowest decision making time for non-overlapping uncertainty scenarios (see Fig. 11). Clearly the machine learning experts had the best professional understanding of probability and recognized the relevant non-overlapping uncertainty tasks to be the easiest, as evidenced by completion in least time (see Fig. 11) and with most confidence. These observations are as expected. From a cognitive perspective, for ML experts this is a case of known probabilities and they feel confident taking up the risk.

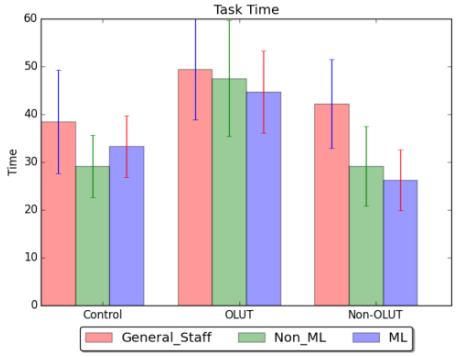

Fig. 11. Average task time in decision making by subject groups.

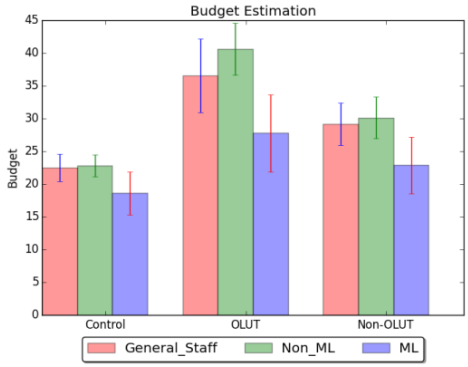

Fig. 12. Average budget estimation of tasks by subject groups.

8.2.3 Non-Machine Learning Experts/Researchers

Non-ML experts seem to fall somewhere between the two subject groups of general staff and machine learning experts. Most non-ML experts find overlapping uncertainty task to be more difficult (see Fig. 10) and report least confidence (see Fig. 9) when making corresponding decisions in it. Also that overlapping uncertainty task is the case which takes them the longest time to complete (see Fig. 11). These observations can be interpreted to be in line with overlapping uncertainty presentation being viewed as ambiguous uncertainty. However, not all of these non-ML experts seem very profound in their perception of probabilistic uncertainty via these visual types. Some of them reported most confident when no uncertainty was presented (see Fig. 9) and coupled with focused budget estimates (see Fig. 12) – this can possibly be interpreted as a tendency to over rely on learning models.

Overall, we can say that only the ML experts probably understood and benefitted from the presentation of uncertainty in these overlapping and non-overlapping type visuals. General staff and non-machine learning experts may have agreed to uncertainty presentation being useful, but were not able to fully benefit from what was being communicated.

8.3 Correlation and Confidence

This study found that revealing of correlations between features and target values did help users more confident in decision making. It was also found that the pattern between correlation and model performance affected user confidence in decision making. For example, when correlation and model performance shared the same trend, users tended to be more confident in their decisions. This was maybe because of the "grounding communication" referred to by psychologists [32]. Because of grounding, confidence in decision making was resolved through a drive towards a mutual understanding or common ground (correlation has the same trend with the performance).

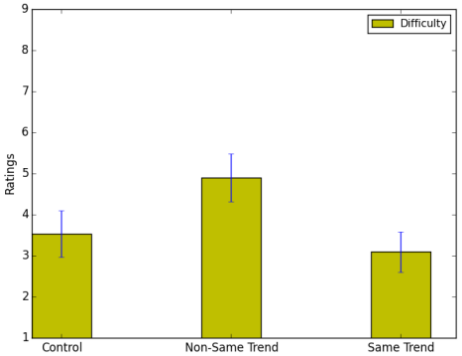

Fig. 13. Average subjective ratings of task difficulty.

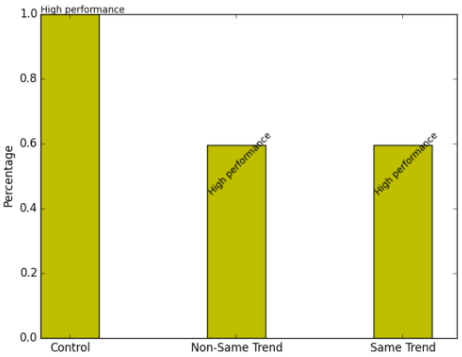

Fig. 14. Choice of models in decision making.

In a separate questionnaire, all subjects were asked to rate the difficulty of tasks. As shown in Fig. 13, revealing correlation data made users confused and felt that decision making was significantly more difficult ($Z=155.5$, $p<.001$) when correlation did not share the same trend with model performance, which decreased the user con-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

fidence. However, when correlation shared the same trend with model performance, users did not feel the increasing of difficulty with the introduction of correlation. Because of the increase in information by revealing correlations, user confidence increased significantly.

By reviewing models participants chose for decisions, most of participants chose high performance models (see Fig. 14). From Fig. 14, it was shown that the revealing of correlation affected the choice of models and decreased the number of participants who chose high performance models. The result suggests that the revealing of correlation information affected both user's decision and confidence.

A group-wise analysis similar to uncertainty-based tasks as presented previously was also conducted and found similar conclusions. For example, ML experts seemed most confident with same trend correlation tasks, while general staff were still most confident in tasks without correlation information.

8.4 Some Limitations

As mentioned earlier, we rely conceptually on "user confidence" ideas, as posited by [33], mainly because the study addressed user confidence and user uncertainty in the context of increasing information that was critically relevant to our investigations. However, there are several developments (in theoretical constructs of "user confidence") like role of individual differences in accuracy of confidence judgements [34], self-consistency model of subjective confidence [35] and collapsing confidence boundary model [36] that must be acknowledged. Here we justify briefly the relevance of our experimental design in the face of some of these developments.

Generally speaking, quantitative studies of decision making have traditionally been based on three key behavioral measures, namely, accuracy, response time (RT) and confidence. Here, confidence is the user's degree of belief, prior to feedback, that the decision reached is correct.

In a typical decision making scenario, once the problem scenario along with supplementary material is presented, several other factors can come into play as well. One such group of factors is individual differences that were investigated by [34]. Differences in experience, motivation, attitudinal predispositions etc. can have an impact on decision making process. However, such differences were minimized, in our case, as we categorized the subjects into professionally skilled groups relevant to predictive task at hand. Next, is the evidence gathering phase, which continues until the user feels comfortable enough to commit to a decision. It is here that studies have shown that the choice certainty (or confidence) is not just influenced by evidence presented (or being gathered) but also by (decision) response time. Greater the time it takes to reach a decision, lower the confidence in that decision [37]. We avoided unlimited RT complications by having several similar decision making scenarios administered repeatedly with a soft encouragement for quick predictive decisions once the basic predictive scenario was understood (in practice sessions). Only items that changed in actual testing procedure were supplementary viewgraph types and corresponding data values.

GSR features derived for the indexing of user confidence are based on 26 participants with four classical classifiers. Despite the reasonable classification accuracies of user confidence, more advanced GSR features and classification models with more participants could be developed to improve the user confidence classification performance.

8.5 End User Design for Interactive Data Analytics

We believe that EUD is the necessary outcome of the IoT condition, because with the advent of IoT, common users will be faced with a situation where everything is digitally connected, giving rise to potential application situations that traditional developers may have never even thought of. Therefore it makes sense to pass on part of the "development" (or personalization/ customization) task to the common user. However, general guidelines for such EUD frameworks are still needed to provide the space in which common user can mold interface functionalities to meet their needs. And it is these general guidelines that we have tried to investigate in the context of interactive data analytics interfaces.

A key functionality of data analytics interface is to support user decision making activity. These interfaces provide this support by presenting on screen ML models learned from historical or streaming data. As we have mentioned in discussion earlier, the key to effective predictive decisions is the user confidence. We explored effects of revealing correlation and uncertainty on user confidence in the data-driven decision making scenario and found that user confidence varied significantly amongst various user groups when different ML models were displayed with/without supplementary information.

This study demonstrated that physiological signals such as GSR features showed significant differences in user confidence levels among tasks. For example, an interesting finding was that less user confidence in decisions was correlated to higher values of GSR features of sum of duration S_d. This was maybe because that GSR signals reflect changes in the skin's ability to conduct electricity and are used to indicate the extent of nerve responses. Less confidence in decisions made users more stressful and users' skin was covered with more sweat, and resulted in the increase of GSR values. GSR features can be used to index confidence levels effectively in decision making.

Some suggestion for EUD for interactive data analytics is that the interface can greatly benefit by including:

- Components which show uncertainty and correlation information. This could help users be more confident in their decisions;
- Information on user confidence levels which allows users make informed decisions.

These components may also be incorporated into the framework of adaptive measureable decision making proposed in [38], therefore introduce confidence levels into data analytics-driven adaptive decision making process. Such user confidence communication in data analytics-driven decision making is more meaningful to both

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

domain users and ML experts and therefore benefits the machine learning transparency. As shown in Fig. 1, the transparent machine learning in turn benefits and plays the key role in the effective EUD for interactive analytics.

9 Conclusions and Future Work

Decision making in urban IoT systems is often driven by machine learning techniques. Human intuitions play significant roles on the success of urban IoT systems. This paper investigated effects of uncertainty and correlation on user confidence in decision making in order to design intuitive user interface for urban IoT systems. A user study found that both uncertainty pattern of model performance, as well as the pattern between correlation and model performance affected user confidence significantly in decision making. Furthermore, the analyses of GSR signals showed that confidence levels in decision making can be effectively indexed with GSR features. These findings have at least two benefits in real-world applications: 1) to design intelligent user interface of decision-related IoT applications. The user interface, which shows user confidence in decision making in real-time, would enhance EUD and help users make informed decisions effectively; 2) to evaluate ML models by measuring user confidence in decision making based on ML output. Supplementary data visualizations contributed to explicability principles while GSR signal indexing added towards correctibility principles. We recommend transparent ML as key to effective EUD for interactive data analytics.

Our future work will focus on analyzing other physiological signals (e.g. blood volume pulse), and behavioral signals (e.g. mouse behavior) of participants for improving confidence classifications during ML-based decision making. Our ultimate goal is to set up a framework of measurable user confidence in predictive decision making in order to dynamically update EUD confidence parameters in urban IoT systems.

Acknowledgment

Authors would like to thank Constant Bridon for valuable contributions to experiments.

References

[1] A. Smirnov, T. Levashova, N. Shilov, and K. Sandkuhl, "Ontology for cyber-physical-social systems self-organisation," in *Proceedings of the 16th Conference of Open Innovations Association (FRUCT16)*, 2014, pp. 101–107.

[2] B. Hoang and S. K. Hawkins, "How will rebooting computing help IoT?," in *Proceedings of the 18th International Conference on Intelligence in Next Generation Networks*, 2015, pp. 121–127.

[3] M. Burnett and T. Kulesza, "End-User Development in Internet of Things: We the People," in *Proceedings of CHI 2015 Workshop on End User Development in the Internet of Things Era*, Sioul, Korea, 2015.

[4] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of Explanatory Debugging to Personalize Interactive Machine Learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces*, 2015, pp. 126–137.

[5] S. K. Sowe, E. Simmon, K. Zettsu, F. de Vaulx, and I. Bojanova, "Cyber-Physical-Human Systems: Putting People in the Loop," *IT Professional*, vol. 18, no. 1, pp. 10–13, Jan. 2016.

[6] F. J. Riggins and S. F. Wamba, "Research Directions on the Adoption, Usage, and Impact of the Internet of Things through the Use of Big Data Analytics," in *2015 48th Hawaii International Conference on System Sciences (HICSS)*, 2015, pp. 1531–1540.

[7] S. Earley, "Cognitive Computing, Analytics, and Personalization," *IT Professional*, vol. 17, no. 4, pp. 12–18, Jul. 2015.

[8] Y. Wang, W. Huang, and H. B.-L. Duh, "Designing and Evaluating a Guiding and Positioning System for Indoor Navigation," in *Proceedings of the 2016 International Conference on Computer Networks and Communication Technology*, Xiamen, China, 2016, pp. 1–9.

[9] F. Chen *et al.*, "Multimodal Behavior and Interaction As Indicators of Cognitive Load," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 4, p. 22:1–22:36, Dec. 2012.

[10] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for Smart Cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, Feb. 2014.

[11] L. He, B. Tang, M. Zhu, B. Lu, and W. Huang, "NetflowVis: A Temporal Visualization System for Netflow Logs Analysis," in *Proceedings of the 13th International Conference on Cooperative Design, Visualization, and Engineering (CDVE2016)*, Sydney, Australia, 2016, pp. 202–209.

[12] J. Zhou, M. A. Khawaja, Z. Li, J. Sun, Y. Wang, and F. Chen, "Making Machine Learning Useable by Revealing Internal States Update — A Transparent Approach," *International Journal of Computational Science and Engineering*, vol. 13, no. 4, pp. 378–389, 2016.

[13] W. Boucsein, *Electrodermal activity*, 2nd ed. Springer, 2012.

[14] C. S. de Souza and C. F. Leitão, *Semiotic Engineering Methods for Scientific Research in HCI*, vol. 2. Morgan&Claypool, 2009.

[15] P. Morgado, N. Sousa, and J. J. Cerqueira, "The impact of stress in decision making in the context of uncertainty," *Journal of Neuroscience Research*, vol. 93, no. 6, pp. 839–847, Jun. 2015.

[16] M. L. Platt and S. A. Huettel, "Risky business: the neuroeconomics of decision making under uncertainty," *Nat Neurosci*, vol. 11, no. 4, pp. 398–403, Apr. 2008.

[17] S. A. Huettel, C. J. Stowe, E. M. Gordon, B. T. Warner, and M. L. Platt, "Neural signatures of economic preferences for risk and ambiguity," *Neuron*, vol. 49, no. 5, pp. 765–775, Mar. 2006.

[18] C. Camerer and M. Weber, "Recent developments in modeling preferences: Uncertainty and ambiguity," *J Risk Uncertainty*, vol. 5, no. 4, pp. 325–370, Oct. 1992.

[19] M. Hsu, "Neural Systems Responding to Degrees of Uncertainty in Human Decision-Making," *Science*, vol. 310, no. 5754, pp. 1680–1683, Dec. 2005.

[20] H. Ibrekk and M. G. Morgan, "Graphical Communication of Uncertain Quantities to Nontechnical People," *Risk Analysis*, vol. 7, no. 4, pp. 519–529, Dec. 1987.

[21] P. M. Allen, J. A. Edwards, F. J. Snyder, K. A. Makinson, and D. M. Hamby, "The effect of cognitive load on decision making with graphically displayed uncertainty information," *Risk Anal.*, vol. 34, no. 8, pp. 1495–1505, Aug. 2014.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2017.2723402, IEEE Transactions on Affective Computing

[22] K. E. Stanovich and M. E. Toplak, "Defining features versus incidental correlates of Type 1 and Type 2 processing," *Mind Soc*, vol. 11, no. 1, pp. 3–13, Jan. 2012.

[23] S. Arshad, Y. Wang, and F. Chen, "Interactive Mouse Stream As Real-Time Indicator of User's Cognitive Load," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2015, pp. 1025–1030.

[24] E. Eyster and G. Weizsacker, "Correlation Neglect in Financial Decision-Making," *DIW Berlin Discussion Paper No. 1104*, 2010.

[25] J. Ye, "Multicriteria decision-making method using the correlation coefficient under single-valued neutrosophic environment," *International Journal of General Systems*, vol. 42, no. 4, pp. 386–394, May 2013.

[26] H. Liao, Z. Xu, X.-J. Zeng, and J. M. Merigó, "Qualitative decision making with correlation coefficients of hesitant fuzzy linguistic term sets," *Knowledge-Based Systems*, vol. 76, pp. 127–138, Mar. 2015.

[27] Z. Li *et al.*, "Water Pipe Condition Assessment: A Hierarchical Beta Process Approach for Sparse Incident Data," *Machine Learning*, vol. 95, no. 1, pp. 11–26, 2014.

[28] W. Wang, Z. Li, Y. Wang, and F. Chen, "Indexing Cognitive Workload Based on Pupillary Response Under Luminance and Emotional Changes," in *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, 2013, pp. 247–256.

[29] D. Kahneman and A. Tversky, "Choices, values, and frames," *American Psychologist*, vol. 39, no. 4, pp. 341–350, 1984.

[30] A. V. Oppenheim and R. W. Schafer, *Discrete-time signal processing*. Upper Saddle River: Pearson, 2010.

[31] J. Healey and R. Picard, "SmartCar: detecting driver stress," in *Proceedings of 15th International Conference on Pattern Recognition 2000*, 2000, vol. 4, pp. 218–221.

[32] H. H. Clark and S. E. Brennan, "Grounding in communication," in *Perspectives on socially shared cognition*, American Psychological Association, 1991, pp. 127–149.

[33] D. K. Peterson and G. F. Pitz, "Confidence, uncertainty, and the use of information," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, no. 1, pp. 85–92, 1988.

[34] G. Pallier *et al.*, "The role of individual differences in the accuracy of confidence judgments," *The Journal of General Psychology*, vol. 129, no. 3, pp. 257–299, Jul. 2002.

[35] A. Koriat, "The self-consistency model of subjective confidence," *Psychol Rev*, vol. 119, no. 1, pp. 80–113, Jan. 2012.

[36] R. Moran, A. R. Teodorescu, and M. Usher, "Post choice information integration as a causal determinant of confidence: Novel data and a computational account," *Cognitive Psychology*, vol. 78, pp. 99–147, May 2015.

[37] R. Kiani, L. Corthell, and M. N. Shadlen, "Choice Certainty Is Informed by Both Evidence and Decision Time," *Neuron*, vol. 84, no. 6, pp. 1329–1342, Dec. 2014.

[38] J. Zhou *et al.*, "Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface," *ACM Transactions on Computer-Human Interaction*, vol. 21, no. 6, p. 33, 2015.

Jianlong Zhou is a Senior Research Scientist of Analytics Group in DATA61, CSIRO, Australia. He is IEEE Senior Member. He got a PhD degree in computer science from the University of Sydney, Australia. His research interests include transparent machine learning, human-computer interaction, cognitive computing, volume visualization, spatial augmented reality and related applications.

Syed Z. Arshad has a PhD in computer science from the University of New South Wales, Australia. He is a member of IEEE, ACM, SIGCHI and SIGKDD. His research interests include cognitive computing, intelligent user interfaces, machine learning and data visualization techniques.

Xiuying Wang is currently a Senior Lecturer, and Associate Director, Multimedia Lab, School of Information Technologies, The University of Sydney, Australia. Her Research interests include biomedical data computing and analysis, biomedical image registration, identification, clustering and segmentation, visual analytics. In conjunction with her students, she has published more than 80 scholarly research papers in the related fields.

Zhidong Li received his PhD degree in computer science from the University of New South Wales, Australia in 2014. He is currently a Senior Engineer of Analytics Group in DATA61, CSIRO, Australia. His research interests include machine learning, computer vision, video analysis, and pattern classification.

(David) Dagan Feng is currently Director (Research), Institute of Biomedical Engineering & Technology, and Academic Director, USYD-SJTU Joint Research Alliance. He has been Head, School of Information Technologies, Faculty of Engineering and Information Technologies and Associate Dean, Faculty of Science, University of Sydney. He has been the Chair Professor, Advisory Professor, Guest Professor, Adjunct Professor or Chief Scientist in different world-known universities and institutes. He is the Founder and Director of the Biomedical & Multimedia Information Technology Research Group at the University of Sydney. He has served as Chairs or Editors of different committees and key journals in the area. Professor Feng has been elected as Fellow of ACS (Australia), HKIE (Hong Kong), IET (UK), IEEE (USA), and Australian Academy of Technological Sciences and Engineering. He received his PhD in Computer Science from the University of California, Los Angeles in 1988.

Fang Chen is a Senior Principal Research Scientist of Analytics Group in DATA61, CSIRO, Australia. She holds a PhD in Signal and Information Processing, an MSc and BSc in Telecommunications and Electronic Systems respectively, and an MBA. Her research interests are behaviour analytics, machine learning, and pattern recognition in human and system performance prediction and evaluation. She has done extensive work on human-machine interaction and cognitive load modelling. She pioneered theoretical framework of measuring cognitive load through multimodal human behaviour, and provided much of empirical evidence on using human behaviour signals, and physiological responses to measure and monitor cognitive load.