Proceedings of the Workshop on

Designing Technologies to Support Human Problem Solving

DTSHPS'18

Lisbon, Portugal on October 1, 2018, in conjunction with the IEEE Symposium on Visual Languages and Human Centric Computing.

Proceedings Editors: Sandra Fan, Narges Mahyar, and Steven Tanimoto, with thanks to the program committee, listed at the workshop website:

www.cs.washington.edu/dtshps2018

Contents, listed alphabetically by first author:

Computer says ‘don’t know’ - interacting visually with incomplete AI models

Alan Blackwell
Computer Laboratory
University of Cambridge
Cambridge, UK
Alan.Blackwell@cl.cam.ac.uk

Luke Church
Computer Laboratory
University of Cambridge
Cambridge, UK
luke@church.name

Ian Hales
Reach Robotics
Bristol, UK
ian@ian-hales.com

Matthew Jones
Judge Business School
University of Cambridge
Cambridge, UK
mrj10@ cam.ac.uk

Richard Jones
Boeing
Bristol, UK
richard.jones16@boeing.com

Matthew Mahmoudi
Department of Sociology
University of Cambridge
Cambridge, UK
mm2134@cam.ac.uk

Mariana Marasoiu
Computer Laboratory
University of Cambridge
Cambridge, UK
mariana.marasoiu@cl.cam.ac.uk

Sallyanne Meakins
Papworth Hospital
NHS Foundation Trust
Papworth Everard, UK
sallyanne.meakins@nhs.net

Detlef Nauck
Applied Research
BT
Ipswich, UK
detlef.nauck@bt.com

Karl Prince
Judge Business School
University of Cambridge
Cambridge, UK
kjp30@cam.ac.uk

Ana Semrov
Computer Laboratory
University of Cambridge
Cambridge, UK
ana.semrov@gmail.com

Alexander Simpson
Computer Laboratory
University of Cambridge
Cambridge, UK
Alexander.Simpson@cl.cam.ac.uk

Martin Spott
HTW Berlin
Berlin, Germany
Martin.Spott@HTW-Berlin.de

Alain Vuylsteke
Papworth Hospital
NHS Foundation Trust
Papworth Everard, UK
a.vuylsteke@nhs.net

Xiaomeng Wang
Computer Laboratory
University of Cambridge
Cambridge, UK
xw337@cl.cam.ac.uk

***Abstract*—This paper presents a design approach to intelligent user interfaces that purposely undermines the perceived "intelligence" of the automated system, with the intention of improving collaborative problem solving. Our goal is for domain experts to maintain a high level of agency, having confidence in their own judgment wherever appropriate, and easily able to question or supplement actions taken by the automated system. We support this approach with four design case studies that incorporate methods from computer vision, natural language processing, data mining, and exploratory visual analytics. Each of the resulting systems has been designed for a specific context of domain expertise. The design guidance derived from these cases relates to the maintenance of uncertainty through visual design cues, encouragement of judgment decisions by expert users, and emphasising the limited evidential status of partial data sets.**

Keywords—intelligent user interfaces, certainty, explanation

I. Introduction

The comedy sketch show *Little Britain* created the catchphrase/meme: ‘computer says no’. Following a long tradition of satirical responses to bureaucracy, this particular meme economically captures the (literal) mindlessness of the supposedly intelligent computer, the frustration of binary decisions in otherwise nuanced human interaction, and the potential for abdication of human agency in information systems. In collaborative problem solving situations, each of these factors presents a serious obstacle to effective collaboration. In our research, we seek design strategies to mitigate those obstacles. We are particularly concerned with innovative design for new technical developments in artificial intelligence and visual interaction, motivated by theoretical perspectives such as mixed initiative interaction [1] and attention investment [2].

We suggest that a more appropriate design stance for such technologies is *computer says don't know*, to be applied in any situation where information is incomplete or alternative courses of action are available - which is to say, *every* situation involving collaborative problem solving. We offer four design case studies that explicitly address specific everyday issues in intelligent system design: incomplete (training) data, ambiguity or uncertainty in inferred models, and availability of human expertise. Each of the four case studies relates to a technical trend in current AI: i) natural language processing, ii) computer vision, iii) ‘big data’ mining, and iv) exploratory visual analytics. Each has resulted in development of an interactive prototype, intended for use by experts in a particular domain: i) international development aid, ii) forensic policing, iii) business decision making, and iv) clinical medicine.

We briefly describe each of these problem domains, and the design strategies taken to support expert problem solving, in the following sections. There are substantial differences between the four design case studies, with regard to the precision of the statistical models, the completeness of the available data, and the complexity of the interactive visual designs. Nevertheless, in this workshop discussion we draw lessons across the four cases to demonstrate consistent design strategies that rectify the unhelpful ‘computer says no’ attitude. The contributions of this work are as follows:

1. Rather than ‘binary’ category judgements (whether yes/no, or larger numbers of logistic classes), we create visualisations that explicitly maintain ambiguity or uncertainty through graphic design cues.

2. Rather than encourage abdication of human agency, we explicitly require the users to make their own judgement decisions through navigating, labelling or constructing interpretive models.

Mariana is a Vice-Chancellor’s Scholar and is supported by an EPSRC industrial CASE studentship co-sponsored by BT. She is also supported by a Qualcomm European Research Studentship in Technology.

The ICUMAP project is part of the Health Foundation’s Insight programme. The Health Foundation is an independent charity committed to bringing about better health and health care for people in the UK.

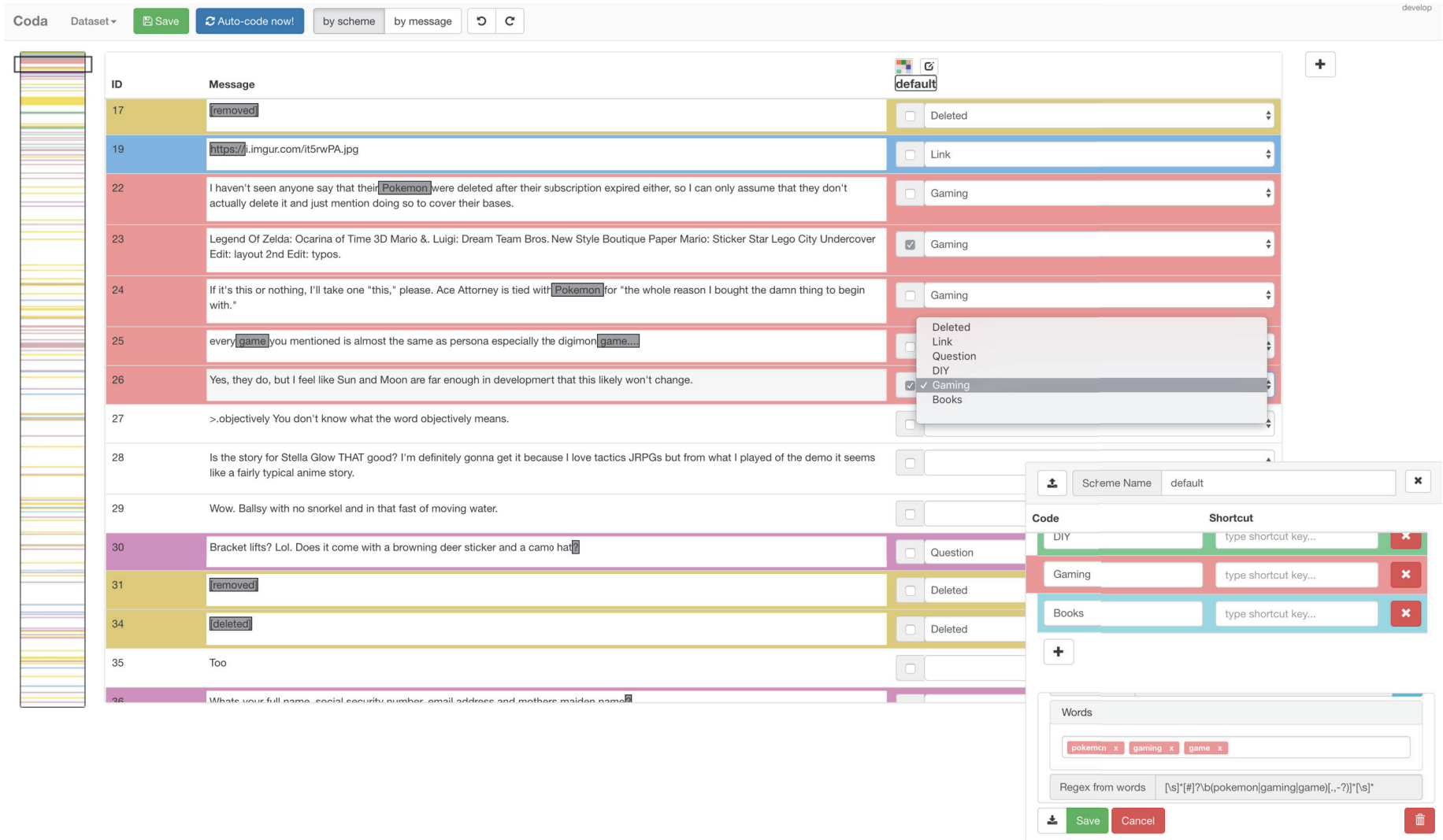

Fig. 1. Coda: each line in the main screen shows a text message, coloured to show how a researcher has categorised it. The available categories (an extensible set) are shown in the inset at lower right. The coloured scrollbar at the left offers an overview of the whole dataset, useful for navigation. The messages are also reorderable to compare categories and confidence in proposed automatic classifications. Since the data Coda is used with is usually sensitive, the data in this screenshot is a sample from the Reddit comment data available on Google BigQuery (https://bigquery.cloud.google.com/table/fh-bigquery:reddit comments)

3. Rather than presenting statistical models as being correct because the data is objective, we draw attention to the ways that the model itself has been created through expectations of usage that frame what can be discussed.

The remainder of the paper discusses each case study in turn, then concludes with a summary of the contribution as demonstrated in those case studies. The studies do represent a wide range of different problem-solving contexts, and we have tried to provide a rich understanding of that context - none of these projects were originally designed to illustrate a simple research theme.

II. Case study 1: Coda

The first case study is a labelling tool designed for an efficient workflow that continually highlights and allows questioning of the categories being constructed, addressing contributions #1 and #2. The application domain relates to the work of Africa's Voices Foundation (AVF), a non-governmental organisation (NGO) whose purpose is to engage with hard-to-reach populations in sub-Saharan Africa through the use of information and communication technologies. AVF works in partnership with local radio stations in remote rural areas, with a focus on conflict regions, low income and low literacy populations. SMS messages from the local population are collected using an SMS gateway and software running on a laptop at the radio station. Community information programmes are accompanied by audience participation surveys, in which listeners are invited to provide a personal perspective on current health issues via SMS. Follow-up questions are sent to listeners who respond to the surveys, requesting demographic information and other survey data. A more in depth description of the methodologies and processes that AVF uses, as well as lessons learnt in designing scalable socio-technical systems for problem solving, is being submitted separately [3].

Analysis of these natural language data-sets is central to the Africa's Voices business model. The "customers" for the analysis are typically humanitarian aid organisations and other NGOs such as United Nations agencies. The business process for Africa's Voices focuses on efficient and reliable coding and analysis of the SMS messages (a difficult process, since the respondents often come from speakers of mixed, low-resource languages), and traceable evidence for communication to the NGO customers. This is the context to our development of an AI-assisted coding tool for use by translators and researchers on the Africa's Voices staff.

The immediate application is a research collaboration between Africa's Voices and UNICEF Somalia, contributing to a programme of drought and famine relief in politically unstable regions of Somalia. In this project, the SMS texts being collected are written in Somali. This introduces further challenges for analysis, as there is no standardised orthography for local Somali speakers, and there are many variant spellings of place-names, as well as diverse cultural attitudes that influence responses to apparently straightforward demographic questions.

Coda is a qualitative coding tool implemented as a Chrome extension (for maximum portability and deployability in diverse environments). The UI supports fast (one-key) decisions that colour-code the data set. As the user works, back-end inference algorithms (currently trivial, but being extended) refine a classifier for semi-automated classification of unseen items. The user can select words in the message to hint to the classifier how they made their decisions.

A colour-coded scrollbar summarises proportion of manually and automated decisions, allowing users to shift between review, correction and refinement to audit and control semi-supervised learning. This colour-coding also serves to explicitly highlight where the computer "doesn't know" how to code a message.

III. Case study 2: ForensicMesh

The second case study proposes an alternative to Computer Vision techniques such as structure-from-motion, by avoiding photorealistic scene rendering in favour of explicitly incomplete models, and highlighting contributions #1 and #3. As more video material becomes available from the extensive usage of Body Worn Cameras (BWCs) in policing, corporate actors have entered the market of building Evidence Management Software (EMS). We anticipate that future EMS systems will use this video to recreate crime scenes and other sites of investigation, as is already done using 3D modelling software and evidential photographs. While there is a plethora of literature and research tackling the adoption of BWCs in policing, work on photogrammetric techniques whereby sites of interest can be reconstructed and verified, and discussions on the implications of Big Data and AI-mediated indicators (see Cheney-Lippold's 'We are data' [4] and Eubanks' 'Automating Inequality' [5]), there appears to be a dearth in research at the intersections of these areas. This has consequences for a range of practices including the emergent practice of predictive policing, counter-terrorism and investigative policing. The ForensicMesh project sought to add value in the process of identifying narratives and storylines as these relate to aggregated video footage from BWCs in policing. With an aim to facilitate a greater space for human judgment in computer-vision aided investigations, as well as for understanding and identifying the distinct and subjective human perspectives of each wearer of BWCs, the project rendered the wearer as a scene element against the backdrop of a static parsimonious scene model, to give analysts access to the human context of data collection.

In high-risk site investigations, BWCs are increasingly used for two possible functions: ongoing monitoring of what are determined to be high-risk sites of potential incidents of terrorism, and captured footage during the investigation of a newly discovered threat or in the aftermath of such a threat. Literature in criminology and sociology has attempted to determine the extent to which the usage of BWCs reduce or increase violence against — and use of force by — the police (see for instance the 2017 report by Barak Ariel [6]). Investigative bodies, such as the NYC Civilian Complaint Review Board, have endorsed the extended usage of BWCs as a step towards greater accountability [7], and there is a general sense that the technology provides objective evidence [8]. According to Wasserman [9], broadly speaking, the position of BWC proponents is commonly summarised into three advantages: 1) 'Video offers unambiguous and objective evidence for all future police-citizen encounters'; 2) 'Video evidence will reduce citizen complains [and] better prove accurate claims and disprove false claims'; 3) 'police and public will behave better knowing that they are being recorded' [9]. While there is plenty to be said for all of these proposed benefits, this project took a point of departure in the first claim.

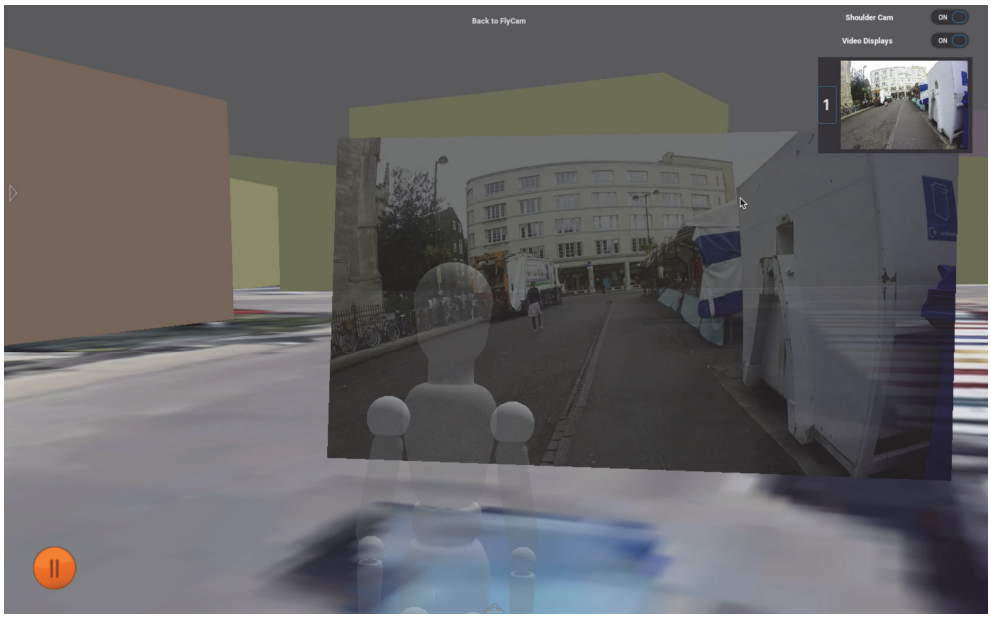

Fig. 2. ForensicMesh: the image shows a parsimoniously rendered citiscape, with only coloured outlines of surrounding buildings. Within the scene, a video plays on a suspended frame, with a human figure in front of it representing the viewpoint of the police officer whose body-worn camera originally captured this video.

Existing research, including by Jones et al. [10], debunk the misconception that BWC footage is "objective", and stress how subjective experiences of incidents shape distinct perspectives of what said footage depicts. Investigative journalists and fact-finders have developed special expertise in creating standards, tools, and methods for using video and image material from official channels as well as publicly available citizen media and other open-source intelligence data. These initiatives (e.g. bellingcat, Forensic Architecture, Amnesty's Digital Verification Corps, Syria Archive etc.) are often attempting to disprove incomplete narratives that may at times be used for political gain (popularly known as misinformation or disinformation). Crucially, they accomplish this by highlighting gaps in the evidence under question, and searching for the "missing link" across a number of different sources. The ForensicMesh project specifically looked at the practices of Forensic Architecture in contemplating environments that would be most revealing of human context and most facilitating of human judgment (and, by extension, doubt).

Forensic Architecture (FA), a visual architecture project group based at Goldsmiths University of London, are widely respected for their contributions to human rights investigations in particular. Their signature immersive 3D reconstructions of sites of violence and crime are particularly worthy of study, as these demonstrate the cutting edge of best practices in establishing visual narratives of events, and filling gaps in data where these exist. Ranging from conflict areas where hospitals or houses may be subject to significant destruction, to murder investigations in Germany, FA use a combination of data-sources including but not limited to: security footage (CCTV), user-generated content (usually in the form of civilian witness footage), images and satellite imagery. Using photogrammetric processing and 3D modelling, FA reconstruct the scene of the particular event in 3D. Videos are hence layered on top of the construction and played in accordance with their sequential timing. Subsequently, linkages between multiple videos and events can be made, and cause and correlation could be established. Virtual sensors such as eye-tracking for subjects in videos can be used to track what event participants are able to see. This approach demonstrates a novel but time-consuming strategy to the reconstructive process. Advancements in computer vision and machine learning, however, mean that the process can be simplified, and — with a critical design and data justice approach — also avoid outsourcing judgment to a machine.

The first aim of this project was to render a 3D reconstruction of the BWC wearer's field of vision, using photogrammetry. This involved the modelling of two primary forms of objects: persistent versus transient objects. Persistent objects include buildings, roads, lights, and signs. Moving, transient objects, on the other hand, include cars, human beings, and animals; these however present a challenge. As transient objects move, their position once outside the field of vision of the BWC wearer can no longer be known and presented with certainty. The predicament presented in such a scenario is evident in the decision between modelling the transient objects in:

1. a predicted position;
2. a fixed position, or;
3. excluding them from the 3D model (once they have left the field of vision). This presents an obstacle regarding the decision to render the scene in 3D or 2D (or both).

In ForensicMesh, we follow FA's approach in the usage of 2D 'video players' embedded within a 3D scene model. A parsimonious 3D scene model of persistent objects (buildings, roads, etc.) can be constructed based on partial information (including images and videos). In our case, we use OpenStreetMap and aerial LIDAR data to generate the parsimonious model. Original footage can then be embedded within a 2D video player to reflect the spatial position and context of the footage, and the wearer of the BWC is represented as a 3D avatar that moves across the scene (Fig 2). The scene can represent multiple wearers with different perspectives of the event. The moving trajectory of each BWC wearer is estimated using a SLAM method [11].

This has several advantages for investigations: First, 2D video players only display what the camera actually captured, leaving no room to display movements and events that were not captured. Second, as the movement of the wearer of the BWC is mapped, a clear timeline of the data-collection event — as well as of the incident itself — emerge, which significantly improves the verification process altogether. Lastly, it emphasizes the temporal and spatial relatedness between multiple BWCs. Where significant gaps are visibly apparent, the analyst is prompted to search for additional sources of evidentiary information beyond the interface.

It is vital to understand the use of tools such as ForensicMesh as a practice that aids — rather than automates — fact-finding processes. 2D video players are deployed where only partial information is captured by BWCs to emphasise new emerging lines of inquiry. In this way, the process of fact-finding becomes a practice of problem solving led by human agents, as gaps prompt the exploration of new leads. These static "gaps" are furthermore an opportunity for the inclusion of open-source intelligence data, including social media, which could shed light on what is not known about the gaps. Explicitly representing these gaps can add value in prompting the investigator to use other resources available to them when the "computer says don't know".

The project set out to develop a photogrammetric tool for video analytics by drawing on best practices in the fields of criminology, forensics, and investigative digital forensics. During the design process, it became apparent that in the development of new and innovative systems for evidence management, it is necessary to build in "uncertainty" by design. Through emphasis on existing or missing connections between BWC footage, ForensicMesh was designed to recenter algorithmically-mediated investigative environments as processes fundamentally of human judgment. This approach not only reiterates the subjective nature of BWC footage, but also demonstrates that ML and computer vision-based technologies can be used outside of regimes that reinforce noxious social biases which are at risk of being algorithmically reproduced.

IV. Case study 3: SelfRaisingData

The third case study is a data visualisation tool for use in the absence of data, highlighting contribution #3. There is an increasing need for business decision making processes to depend on analysing large quantities of data. However, not all the data is easily available or even collected when questions and hypotheses arise, nor is there much time in the fast paced context in which business managers operate to sit down with an analyst and explain and detail the high level question into deliverables. Data analysts have only a short time after they receive an analysis request to clarify the business manager's question. They rely on their expert knowledge of the business domain and of the organisational context to anticipate the (implicit) needs of the business.

The focus of this project was building a data visualisation tool that would support remote collaboration between data analysts and business managers requesting data analyses and reports. Through several interviews with data analysts working for BT, we identified a number of challenges in their existing workflows, from difficulties of data extraction to the importance of careful communication of results to non-experts. A more in depth description of the research methodology and of the tool created can be found in Mărășoiu et al. [12]; here we describe the part of the analytical process that we chose to design a solution for, and briefly describe the technical artefact created, emphasizing the three design strategies discussed in Section 1.

In this project, we chose to focus on the hypothesis clarification and refinement part of the analytical process - the conversation between the analyst and the requester, where the former aims to better understand the question being asked by the latter.

Some of the analysts we interviewed work on many small data analysis projects, often at the same time. High level business questions such as "why is [this region] not as good as everybody else, what is happening there" are the typical starting point for such a project. But before the analysts can get started on finding and extracting the needed data from the company's databases and data silos, they need to both have a better understanding of what the requester needs, and to turn the high level question into actionable steps and outcomes. They need to add parameters to the original hypothesis, and "fill in all the details", by asking more questions about e.g. which aspects of the region they should look into, what "everybody else" means, what "not as good" means. The data analysts we interviewed had these kinds of conversations primarily through phone calls and sometimes via emails.

Our system, SelfRaisingData, allows sketching and modifying data visualisations in order to support these remote conversations structured by analytic hypotheses.

Since the analysts we interviewed had such conversations primarily through phone calls, the typical scenario would be that both the analyst and the manager would be working on the same visualisation document on their own computers, whilst on the phone with each other. The visualisation has the role of being an external representation of the analyst's understanding of the manager's question, with the manager being able to comment and point out where their understanding of the question diverges. As such, the analyst can create and edit the visualisation to reflect their understanding, whereas the manger can only annotate it.

Fig. 3 illustrates the system. The choice of visualisation comes from the domain in which the analysts we interviewed worked in, as a large part of the type of data they work with is timeseries data. The central area of the system represents a timeseries visualisation of synthetic data. The data is generated from additively composing a set of parameterisable functions added from a tool panel to obtain a trend line. Each component function can be parameterised independently in a function editor by dragging the value handles of its properties on axes corresponding directly to the axes of the final visualisation. The user can add, remove, and modify each component function.

Whilst the resulting composed function could be displayed as a line chart, this visualisation style can result in users fixating on manipulating the parameters of the composing functions in order to achieve a smooth line. Instead, we add noise to the trend line by 1) transforming the continuous function into a discrete set of points by sampling N equally-spaced time coordinates (the X axis) and 2) sampling the value coordinates (the Y axis) from standard distributions having the value of the trend line function as their mean and a constant fixed variance. To further suggest sketchiness and imprecision, we represent each individual point as a hand drawn cross.

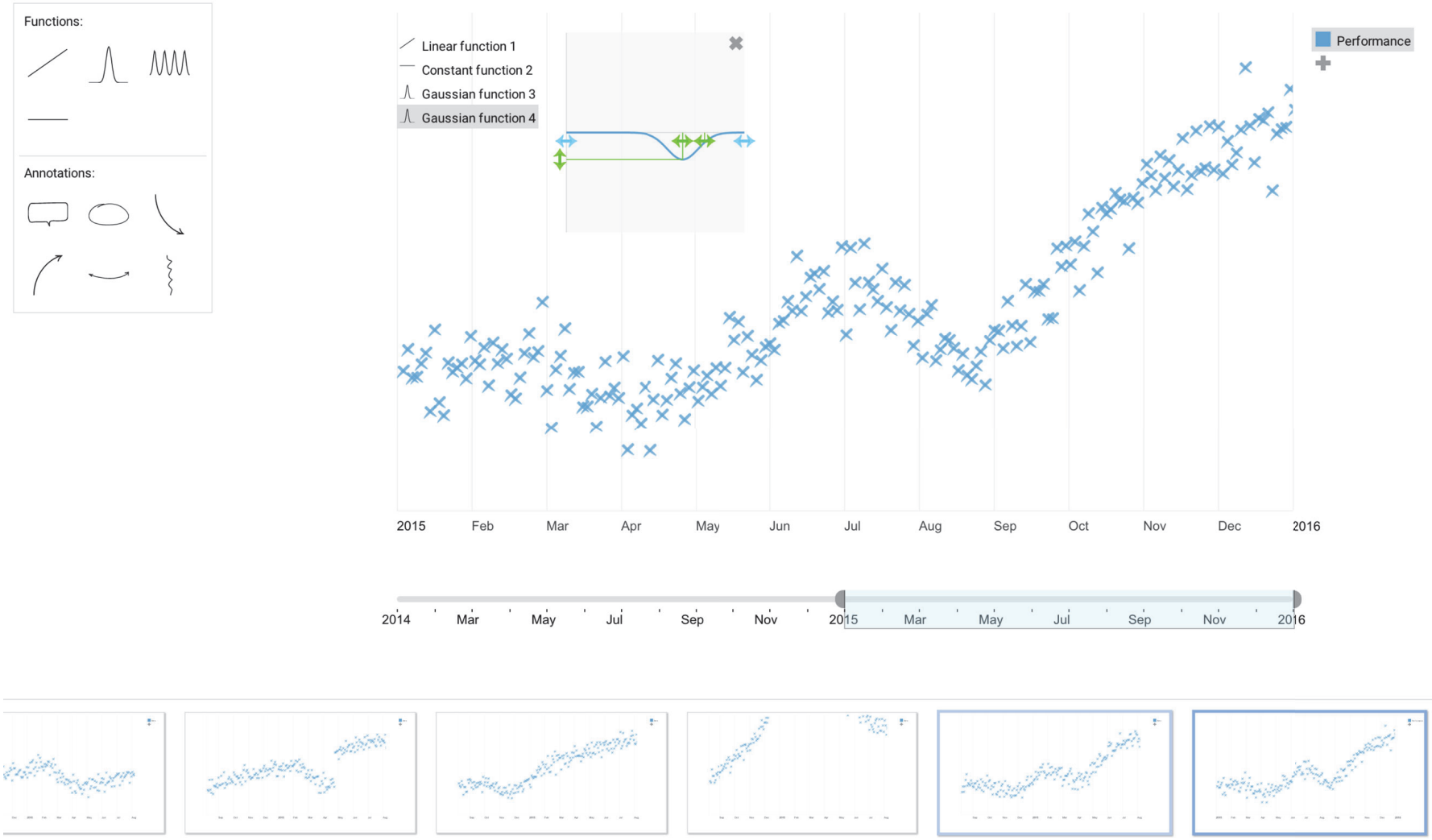

Fig. 3. SelfRaisingData: the time series chart contains synthetic data generated around a trendline through function composition of basic functions (linear, constant, squared exponential and periodic). Each component function can be independently modified in an interactive function editor. The history panel at the bottom contains snapshots of the evolution of the visualisation, allowing the analyst to revert to a previous version and explore alternative hypotheses.

A panel at the bottom of the screen records the history of the visualisation in conceptual sequences of steps (e.g. adding, removing, and editing a different component function, adding a new timeseries). The users can change their mind at any point, having the ability to load any of the past versions of the visualisation and create a new development branch by editing the older version.

Our design goal in this project was to support a conversation about data in the absence of real data through sketching. In this case study, the computer "doesn't know" anything, but provides support for collaboration and statistical problem specification. The visualisation system is a conversational aid that frames the conversation between two people. For example, treating the visualisation as a composition of independent elements (e.g. trends, periodicity, plateaus, dips and peaks) is a deliberate design choice. It adds hypothesis semantics to the visualisation as each component function acquires an individual meaning (e.g. the March-June performance plateau, the 8th of October sales drop). Further, independently manipulated component functions are still available for discussion even after being composed with other functions. The always-visible list of component functions also draws attention to the way that the sketch has been constructed.

Since the data visualised is synthetic and created by the user, emphasizing ambiguity is also relevant. We achieve this through the noisiness of the scatter plot, which also allows for imprecision when adjusting the parameters of the component functions. This means that sketching can be done quicker, as (spending time to achieve) precision is actively discouraged. Vagueness is further encouraged by representing each point as a hand drawn cross and removing any numbering from the function editor panel. The visualisation is a rough sketch of how the real data might look like.

V. Case study 4: ICUMAP

Our final case study is an interactive visualisation that was created to support clinical judgments in an intensive care unit (ICU) through reuse of electronic health record (EHR) data, highlighting contributions #2 and #3. During treatment in an ICU, large amounts of data are collected for each patient, including both nursing observations and automated data acquisition from monitoring instruments (e.g. blood pressure and pulse) at the bedside. Subsets of this data are collated from hospitals around the UK by the national intensive care registry (ICNARC), which uses it to calculate statistical measures of patient condition for comparison to treatment outcome. However, our research with clinicians across multiple hospitals suggested that these measures have low predictive power and are never used directly to guide treatment, perform triage, alert potential emergencies, or otherwise guide clinical judgment.

We had access to 10 years of data, covering the treatment of 20,000 patients, from an ICU specialising in cardiothoracic surgery (e.g. arterial bypass grafts, heart valve replacements and heart transplants). While standardised statistics such as ICNARC are compiled based on a small number of physiological measures at admission time, we were able to pay attention to how the patient's condition changes during the time they are in the unit. In particular, our clinical collaborators wanted to know when a patient's condition changes in a way that is likely to have adverse outcomes - expressed to the design team as a 'traffic light' indication.

The system we designed, ICUMAP, is a dimension-reduced visualization (Fig. 4), in which a variant of t-SNE [14] is used to construct a reference 'map' of regions in which intensive care patients are 'similar' within a multi-dimensional space of variables monitored during their treatment. The condition of each patient is mapped to a new location at 6-hour intervals, and these points are joined to form a trajectory. Early experiments confirmed that some regions in the t-SNE cluster map were associated with high mortality, meaning that the 'traffic light' goal could apparently be expressed as places where a patient appeared to be 'moving toward' a high mortality region in t-SNE space. However, this is not a mathematically well-formed question. t-SNE can be interpreted as a projection of a multi-dimensional space, but the projection is not necessarily monotonic in any dimension. Our experiments confirmed warnings of Wattenberg et al [15], that possible "tendencies" were simply a tangle of overlaid random lines.

We therefore modified the t-SNE algorithm, adding two new constraints to the distance function. The first was to penalise long distances between successive measurements for the same patient, creating temporal locality as a basis for thinking about trajectory. The second was to promote proximity for measurements taken at the point where a patient had died, meaning that these were more likely to cluster together, with the result that mortality would correspond to particular "places" within the optimised layout. The third was to include the surgery that the patient had undergone as a strongly weighted factor, meaning that cases tended to be grouped according to procedure, corresponding to natural classifications used by clinicians.

We optimised the visual rendering to convey local detail of individual trajectories, while also offering a distribution overview of thousands of these. This involved manipulation of hue gradients, line widths, and alpha (transparency) values so that each trajectory could be viewed as a progression from condition at admission (blue at the start of the line) to either mortality or discharge (red or green). Within the clusters of different surgery types, those types that are more risky can be identified by greater density of red trajectories.

This design reflects a Bayesian approach to clinical decision making. Rather than claiming statistical likelihood of treatment outcomes, we focused on assisting clinicians by improving ease of access to relevant prior cases among the thousands they might consider, and making this information more readily available as a counter to the usual heuristic biases in clinical judgment. We therefore focused on selecting a small number of comparator cases, sufficiently similar to suggest relevance to a patient currently under consideration, but presented within an interpretive metaphor that would facilitate reflection on the current case, while not over-determining the conclusions that might be drawn.

ICUMAP includes many features for interaction with the data archive, always aiming both to provide users with statistical overviews while comparing and contrast a patient currently being treated with cases from the historical records. When mousing over the map, the trajectories for individual patients are highlighted. The measured values for that patient at this time are shown on histograms showing overall distributions, so users can see at a glance how typical this patient is. When one histogram is selected, a mask over the map visually fades areas where the value of this variable is low.

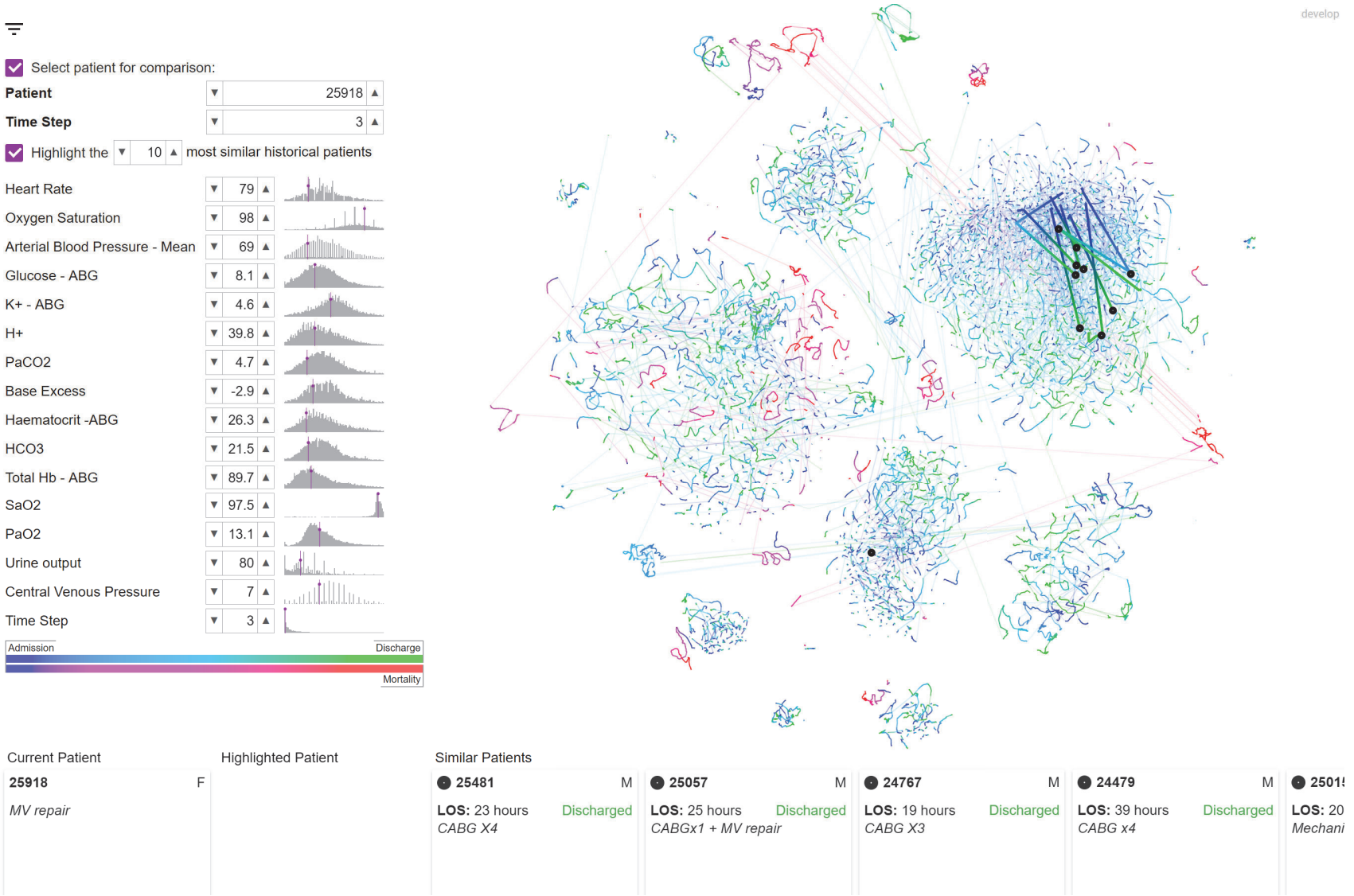

Fig. 4. ICUMAP: the "cloud" shows trajectories of thousands of individual patients within a dimension-reduced multivariate space. Clusters correspond to surgical procedures. At the left, histograms show distribution of values for each of the principal variables, across the whole dataset. At the bottom, patient records are shown for 10 patients whose condition was most similar to the one currently selected by the user.

These features allow clinicians to review prior likelihood, variability, value exceptions, and other broad statistical properties of the large database. However, it is necessary to keep in mind that the t-SNE 'landscape' does not necessarily provide any continuity with respect to a single variable – the variation in masks and intensity is therefore rendered with large tiles to reveal this blotchy variability, and discourage misleading interpretation.

In this final case study, we explicitly evaluated the extent to which our design approach is understood and appreciated by domain experts. This involved a focus group with expert staff, and controlled task observation with students.

A. Focus Group Evaluation

We invited eight clinicians and a hospital information analyst to a focus group workshop, at which we demonstrated the functionality of ICUMAP, explaining each aspect of system functionality before pausing for comments and discussion.

We report two primary topics of discussion that arose. The first was the question of which data in the EHR database has the most clinical value for predicting or modifying treatment outcomes. There were diverse opinions, with some senior clinicians strongly advocating use of particular measurements (some of which are not presently recorded in the EHR at all). The second topic focused on the main theme of this paper, which was skepticism regarding the 'sales' message of predictive analytics. Many participants had encountered products that claimed to deliver predictive functionality through multivariate data mining. Clinicians were skeptical that such prediction was possible. Their view was that single variables reflect critical aspects of patient condition, and that multivariate analysis (and hence dimension reduction visualisations) does not significantly add to clinical judgment.

Nevertheless, our central design strategy, drawing clinical attention to a small number of previous cases similar to the current patient, appeared to be welcomed. Scepticism about the value of predictive analytics was directed at other systems (or speculation about what our system might be), while there was productive conversation about the identified similar patients.

B. Controlled experiment evaluation

We recruited six participants to evaluate ICUMAP in a controlled task. Three were clinical professionals (two medical students nearing the end of their studies, and one registrar intensivist), and three students from non-medical (engineering, physics, computer science) backgrounds. A predefined data set was loaded, and each participant worked through the same series of interpretive tasks. At the start of each task, the participant was asked an interpretive question without prompting them about ICUMAP functionality. If they had not recognized the expected functionality, the relevant system function would be explained (using a predefined text) before proceeding.

In order to compare interpretation of the t-SNE cluster visualization to more conventional statistical visualisations, participants were first shown a screen with only the histogram distributions. The participant was asked questions regarding their interpretation of these historical distributions, and then shown values for a small number of individual test patients drawn from the database to represent distinctive types, before being offered the opportunity to compare these individual patients to the overall population. The clinical participants were asked how they would interpret the condition of each test patient in their clinical judgment. After making their interpretations, the clinical participants were additionally asked to report how confident they were.

The t-SNE cloud visualization was then revealed, with the explanation that this represented change over time for the same measurements. Participants were asked for their unprompted interpretation of design elements. If they did not volunteer key aspects (time-courses, proximity, mortality), these were explained. Features were tested in turn, each time offering an opportunity for the participant to make their own interpretation before the design was explained. Finally, a small number of patients were selected, each chosen to represent a particular type of surgery or outcome. Participants were asked for their interpretation of likely treatment outcomes, taking into account other patients automatically highlighted as 'similar'. At each point where participants offered an interpretation, they were asked to quantify their level of confidence in that judgment.

We found that while all technical participants recognised histograms as describing statistical distribution, two medical participants initially *misinterpreted* histograms as representing change of a measure over time (the existing EHR system presents a patient overview with prominent time-series graphs). Once they understood the principle, they were able to use the data for assessments of a single patient. However, they relied on *prior expectation* of typical values (i.e. a value range learned during their studies) for initial assessment. Where they had less prior knowledge, they paid more attention to the plotted position of a value within the overall distribution. They used a time step control to explore progression and discuss changes in the patient's condition over time, for example a crisis at one time step. They expressed more confidence in judgments when exploring this historical data.

Technical participants immediately recognized that the ICUMAP visualization was a dimension-reduced view of multivariate data. None of the medical participants recognized this, and found the visual complexity overwhelming. After using the mouse to explore trajectories, they were able to identify properties of the visualization, although one remained uncomfortable throughout the session. All understood that the lines represented the trajectory for a patient, and that red and green reflected mortality. None recognized the basic principle of similar points being near each other. Two of the three recognised without prompting that clusters reflected type of surgery.

Medical participants, in interpreting the overall structure of the cloud, tended to make comparisons between clusters. One expected (incorrectly) that larger clusters might reflect wider distribution of values, while others observed correctly that cluster size related to the number of patients in that cluster.

The key principle of selecting and plotting a group of similar patients was recognized without prompting by all medical participants. When asked to make judgments based on this visualization, they did, as intended, immediately start to make comparisons, for example by starting to talk about relative length of stay, that they did not do when considering histograms alone. A major concern of ours was to avoid over-interpretation of the similarity as predictive data. None of the participants expressed a confidence of 100%, with most

judgments being in the 50% to 80% reflecting a suitable degree of caution. One of the test patients, having healthy values and routine surgery, where all similar cases had been discharged successfully, led a participant to give a 90% assessment that this patient would also be discharged. These findings are encouraging, however, we noted a trend that the confidence judgments tended to increase over the course of the experiment, suggesting that growing familiarity with a tool may still lead to errors of the kind that we wish to avoid through our design philosophy.

VI. Discussion

In this section, we summarise the design strategies that have been applied in these systems, relating them to the three broad contributions outlined in the introduction to the paper.

Firstly, we suggest using graphic design cues to maintain ambiguity or uncertainty in presenting inferred information to users. This is a corrective to the increasing tendency in many machine learning systems to present the categorical output variables of logistic regression as simplistic either/or alternatives, replicating the errors that were systematically identified in the seminal work by Bowker and Star [16]. In Coda, we use desaturated colours to contrast automated suggestions with human-assigned labels, allowing natural interpretations such as complete desaturation (white) being equivalent to no judgment at all, while near-full saturation indicates that the system has identified duplicates that are safely amenable to trivial automation. In ForensicMesh, we eschew photo-realistic scene rendering in order to remind viewers that a geometric model is based only on a persistent coordinate system, not fully-observed state.

Secondly, we suggest that users should be explicitly required to make their own judgement decisions. In a labelling system such as Coda this is trivially true. However, we should recall that most AI systems intentionally hide the labelling phase (usually done offline, via a different interface, and prior to system operation). In ICUMAP, we do not directly present statistical regression on patient condition as a basis for decision making, instead emphasising the clinician's responsibility to retrieve and consider other cases - based on the particularity of clinical interventions and patient case histories.

Thirdly, we suggest that systems draw attention to the ways that the model itself has been created through human processes, reminding users that these processes anticipate the ways the model can be used. The most extreme is SelfRaisingData, in which users are invited to completely "fabricate" data to reflect their ideas about the model. In ForensicMesh, we insert an "observer" into the scene, to emphasise that BWC video is not objective, but reflects the viewpoint of the person who was wearing the camera. In ICUMAP, the use of a query / recommendation interaction paradigm means that the "model" is transient, presented only as a byproduct of the user's brushing over a cloud of patient journeys, or over distributions of measurement values.

Each of these design strategies has potential for use in other collaborative problem-solving settings, and we look forward to workshop discussion considering analogies to other intelligent interaction scenarios.

VII. Conclusion

Although "computer says no" was introduced as a comedy trope, the extension of algorithmic decision making throughout society has become tragic, as when a British man was denied an ambulance because the triage algorithm determined that his case was not serious, despite the fact that he was in agony, correctly diagnosed his own condition, and subsequently died. Cheney-Lippold [4] quotes the operator "We cannot override this, and although there are paramedics in the control room for us to ask, I would not think the system would come up with the wrong answer"

Our four case studies all involve use of data in mission-critical or safety-critical settings. Enhancing reliability of data analysis in such settings is obviously an important research goal for data science and AI. However, at present, these are domains where expert human judgment is respected and human experts take responsibility (and liability) for their interpretations and decisions. This paper has considered a number of visual language design approaches through which expert responsibility can be maintained, with 'intelligent' analysis focused on making the necessary data salient and easily available for human judgment, rather than taking automated decisions.

References

[1] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 159–166.

[2] A. F. Blackwell, "First steps in programming: A rationale for attention investment models," *Proceedings - IEEE 2002 Symposia on Human Centric Computing Languages and Environments, HCC 2002*, pp. 2–10, 2002.

[3] L. Church, R. Zágoni, A. Simpson, S. Srinivasan, and A. Blackwell, "Building socio-technical systems for representing citizens voices in humanitarian interventions," submitted to *Designing Technologies to Support Human Problem Solving - A Workshop in Conjunction with VL/HCC 2018*, Lisbon, Portugal.

[4] J. Cheney-Lippold, *We Are Data: Algorithms and The Making of Our Digital Selves*. NYU Press, 2017.

[5] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[6] B. Ariel, A. Sutherland, D. Henstock, J. Young, and G. Sosinski, "The Deterrence Spectrum: Explaining Why Police Body-Worn Cameras 'Work' or 'Backfire' in Aggressive Police–Public Encounters," *Policing: A Journal of Policy and Practice*, vol. 12, no. 1, pp. 6-26, 2017.

[7] Civilian Complaint Review Board, "Civilian Complaint Review Board Issues 2017 Annual Report," New York City, 2018.

[8] D. K. Bakardjiev, "Officer Body-Worn Cameras - Capturing Objective Evidence with Quality Technology and Focused Policies," *Jurimetrics*, vol. 56, no. 1, pp. 79–112, 2015.

[9] H. M. Wasserman, "Recording of and by Police: The Good, the Bad, and the Ugly," *J. Gender Race & Just.*, vol. 20, p. 543, 2017.

[10] K. A. Jones, W. E. Crozier, and D. Strange, "Believing is Seeing: Biased Viewing of Body-Worn Camera Footage," *J. Appl. Res. Mem. Cogn.*, vol. 6, no. 4, pp. 460–474, Dec. 2017.

[11] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 3, pp. 611–625, 2018.

[12] M. Mărășoiu, A. Blackwell, A. Sarkar, and M. Spott, "Clarifying hypotheses by sketching data," in *EuroVis 2016 - Short Papers*, 2016.

[13] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J. D. Fekete, "Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2769–2778, 2012.

[14] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, 2016.

[16] G. C. Bowker and S. L. Star, "Sorting Things Out: Classification and Its Consequences". MIT Press, 2000.