Using Machine Learning to Infer Reasoning Provenance from User Interaction Log Data: Based on the Data/Frame Theory of Sensemaking

Neesha Kodagoda[1], **Sheila Pontis**[2], **Donal Simmie**[3], **Simon Attfield**[1], **BL William Wong**[1], **Ann Blandford**[2], and **Chris Hankin**[3]
[1]Middlesex University, UK, [2]University College London, UK, [3]Imperial College London, UK

The reconstruction of analysts' reasoning processes (*reasoning provenance*) during complex sensemaking tasks can support reflection and decision making. One potential approach to such reconstruction is to automatically infer reasoning from low-level user interaction logs. We explore a novel method for doing this using machine learning. Two user-studies were conducted in which participants performed similar intelligence analysis tasks. In one study, participants used a standard web browser and word processor; in the other they used a system called INVISQUE (INteractive Visual Search and QUery Environment). Interaction logs were manually coded for cognitive actions based on captured think-aloud protocol and post-task interviews, using Klein, Phillips, Rall, & Pelusos' Data/Frame model of sensemaking as a conceptual framework. This analysis was then used to train an *Interaction Frame Mapper* which employed multiple machine learning models to learn relationships between the interaction logs and the codings. Our results show that, for one study at least, classification accuracy was significantly better than chance and compared reasonably to a reported manual provenance reconstruction method. We discuss our results in terms of variations in feature sets from the two studies and what this means for the development of the method for provenance capture and the evaluation of sensemaking systems.

Keywords: analytic provenance, reasoning provenance, data provenance, sensemaking, data/frame model, interaction frame mapper, machine learning.

INTRODUCTION

Reconstructing analysts' reasoning processes during complex sensemaking tasks has the potential to provide insights about those processes. This may shed light on how complex investigations are conducted. Such a record can be a resource for '*reflection-in-action*' (Schon, 1983) during an investigation, for planning or reframing of objectives and scope, or it may be a resource for *'reflection-on-action'* after the event, for audit, training or evaluating outcomes. This may be particularly important in task domains where accountability for outcomes is important, such as in intelligence analysis, and investigations by the police and lawyers.

Complex human cognitive activities can be represented at different levels of description. Interaction logs represent a history of an analyst's interactions with a system and fall into

a category that has been referred to as *analytic provenance* (Roberts et al., 2014). While logs can be captured relatively easily, they are low-level and provide no representation of the analyst's reasoning. One possibility might be to use this information to reconstruct what has been referred to as *reasoning provenance* (Roberts et al., 2014) i.e. a trace of the analysts reasoning process. The challenge is that this information is usually tacit and '*in the head*' of the analyst. However, just as we might abductively infer intent from observations of human action based on an understanding of how intentions can be expressed through action, so might we potentially recover *reasoning provenance* from *analytic provenance*. In this paper we report an experiment in recovering *reasoning provenance* from *analytic provenance* via a cognitive model of sensemaking. Table 1 presents definitions of key terms used throughout the paper.

Table 1

Definition of terms.

INteractive Visual Search and QUery Environment (INVISQUE)	An interactive visual reasoning system which supports search and the freeform manipulation of information on an infinite canvas.
Data provenance	A trace of data from its origins including its movement between databases.
Analytic Provenance	A history of an analyst's interactions with a system.
Reasoning Provenance	A history of the thinking, reasoning and decision processes underpinning analytic steps.
Frame	A representation (in the mind of an analyst) which offers an account, perspective, viewpoint or hypothesis of a situation.
Data/Frame Model (DFM)	A model of sensemaking which presents sensemaking as a process of fitting data info a frame or fitting a frame around the data.
Interaction Frame Mapper (IFM)	A machine learning algorithm used to map functions between descriptions of interactions and associated sensemaking actions.
Frame manipulation action (FMA)	A sensemaking 'action' characterised by the Data/Frame Model.
Support Vector Machines (SVM)	A discriminative classifier defined by a separating hyperplane. When given labelled training data, the algorithm outputs optimal hyperplane which categorizes new examples.
Random Forest (RF)	A classifier that consists of many decision trees and outputs the class that is the mode of the class's output by individual trees.
Hidden Markov Model (HMM)	A tool for representing probability distributions over sequences of observations.
Radial Basis Functions (RBF)	Approximate multivariable functions by linear combination of terms based on a single univariate function.

We developed an *Interaction Frame Mapper* (IFM) to perform a mapping function between descriptions of interface interactions *(analytic provenance)* and descriptions of associated reasoning *(reasoning provenance)*. The mapping function used three different supervised machine learning models which were evaluated independently in terms of performance. Training data were provided by two user-studies in which participants performed investigation tasks. In one study, participants used a standard web browser and word processor; in the other they used a system called INVISQUE (*INteractive Visual Search and QUery Environment*). INVISQUE is an interactive visual reasoning system which supports search and the freeform manipulation of information on an infinite canvas. In both studies, we captured low-level user-interaction logs, think-aloud protocols and post-task interviews. These then provided a basis for inferring reasoning provenance traces by coding the protocols (broadly) in terms of Klein, Phillips, Rall &

Peluso's (2007) Data/Frame Model (DFM) of sensemaking. This model generalises across domains of sensemaking and, consequently, provided a generalizable language for describing *reasoning provenance*.

In the following, we begin by discussing related research followed by a description of the two user-studies that provided our training data and of the training process. We then report the performance of the mapping function in relation to these two different studies and the three machine learning models. Finally, we discuss our findings and future work.

RELATED WORK

Analytic Provenance and Reasoning Provenance

Reconstructing *reasoning provenance* from interaction with a visualization has been described by North, et al., (2011) as "...understanding a user's reasoning process through the study of their interactions ...". They propose a five-stage model for reconstructing *reasoning provenance* information:

1. **Perceive:** Understand how data is presented to the user (as a resource for disambiguating interaction data);
2. **Capture:** Record the sequence of interactions;
3. **Encode:** Describe the captured provenance in a predefined format;
4. **Recover:** Make sense of the history of actions performed by a user i.e. recover reasoning;
5. **Reuse:** Apply the provenance of previous analyses to new data and/or domains.

In these terms, the focus of the current paper is on stage 4, the recovery process; that is, how inferences might be drawn from interaction data.

Roberts, et al., (2014) classify provenance information into: *data*, *analysis* and *reasoning* provenance. *Data provenance* concerns the path between the data source and the system. This traces data from its origins including its movement between databases; *analytic provenance* refers to the sequence of actions taken in producing an analytic product; and *reasoning provenance* refers to the history of reasoning during the same process. Stages 3 and 4 in North et al. (2011) effectively make the distinction between the record of the analysis actions and its associated thought processes made by Roberts et al. (2014). *Analytic provenance* is about the interaction with the system, tracing interaction histories through system states, and *reasoning provenance* is about associated thought processes guiding and resulting from interaction in combination with the analyst's background knowledge and assumptions. In these terms, we are seeking to recover reasoning provenance information from information about the analysis. We do this by mapping from interactions to reasoning activities via an established model, Klein, Phillips, Rall & Peluso's (2007) DFM.

An example of manual recovery of *reasoning provenance* was reported by Dou, et al., (2009). They described a process in which trained financial analysts used interaction logs

of an analysis of suspicious wire transfers using a visual analytics tool to code for reasoning processes. The *reasoning provenance* trace included descriptions of findings, strategies and methods. For example, a common strategy was to look for gaps in an operational timeline. Interpreting these gaps as the points where findings were established, they then worked backwards to identify strategy and method. They report that the coders had difficulty identifying methods used for the investigation if these were based on visual patterns. They suggest that their approach works best for highly interactive visual systems. Being the only directly comparable approach to our own, we return to the results of Dou et al., (2009) at the end of the paper in the section *Class-based F1 Measure Analysis*.

Gotz and Zhou (2008b) report the use of combined manual and automatic capture of both semantic and comprehensive records of user activity with minimal user involvement to generate what they refer to as *insight provenance*—a record of the process and rationale by which an insight was derived during a visual analytics task. They propose capturing lower-level interaction logs, and user analytic behaviour at multiple levels of granularity based on the semantic richness of the activity. Their approach was motivated by Activity Theory (Nardi, 1995) and their prior analysts' behaviour (Gotz & Zhou, 2008a). They define tiers from rich to poor semantics: *task, sub-task, action* and *events*. The focus is on *action*, which falls between high-level user goals and low-level user interactions. *Action* refers to the analytic steps performed by the analyst, while *events* are the low-level user-interactions. After review of several visual analytics systems a set of actions are identified and characterised by type, intent and parameters of each user-action. These are further developed into three general classes of action: *exploration actions*, *insight actions* and *meta actions*. The initial findings suggest that the approach of capturing *insight provenance* was promising.

Chen, Qian, Woodbury, Dill, & Shaw (2014) used a parametric symbolic model (dependency graph) to represent the provenance of an analysis. As the user interacts with a Visual Analytics tool, the symbolic model is parsed automatically from the interactions. Although Chen's system provided automated capture of *analytic provenance*, the recovery of *reasoning provenance* was left to the analyst by browsing the dependency graph and visualisation history. In contrast to Gotz and Zhou (2008b) and Chen, Qian, Woodbury, Dill, & Shaw (2014) our aim is to recover *reasoning provenance* from *analytic provenance* without the need for manual intervention based on a learned mapping Sensemaking.

Our language of *reasoning provenance* is based on Klein, Phillips, Rall & Peluso's (2007) DFM of sensemaking (Figure 1). Sensemaking has been described as the process of finding meaning from information (Weick, 1995) and as "the deliberate effort to understand events". It is a process through which people draw inferences, make predictions and generally gain knowledge from data (Klein, Phillips, Rall & Peluso, 2007). Most accounts of sensemaking describe it as a bi-directional interplay between bottom-up and top-down cognitive processing which seeks to reconcile data on the one hand with emerging representations which account for that data on the other (Klein, Phillips, Rall & Peluso, 2007; Starbuck & Milliken, 1988; Weick, 1995).

Of these accounts, Klein, Phillips, Rall & Peluso's (2007) DFM of sensemaking delineates underlying cognitive processes most clearly. The DFM distinguishes two kinds of entity: data and frame. Data are aspects of a prevailing situation that a sensemaker experiences as they interact with it. A frame is a representation which offers an account, or hypothesis of that situation.

The DFM describes seven types of sensemaking process or cognitive actions (Figure 1):

- **Connecting data to frame:** Some data are understood within the context of a frame, or interpretation. The frame may stand as an initial and possibly tentative understanding of a given situation.
- **Elaborating the frame:** Searching for data that might extend understanding of the situation. The frame defines what counts as data.
- **Questioning the frame:** Questioning the validity of a frame given data which violates expectations that the frame sets up.
- **Re-framing:** Dropping an existing interpretation of a situation in favour of a new one.
- **Preserving the frame:** Explaining away inconsistencies between data and frame.
- **Comparing frames:** Considering the best candidate out of a number of interpretations.
- **Seeking the frame:** Explicitly searching for a new frame.

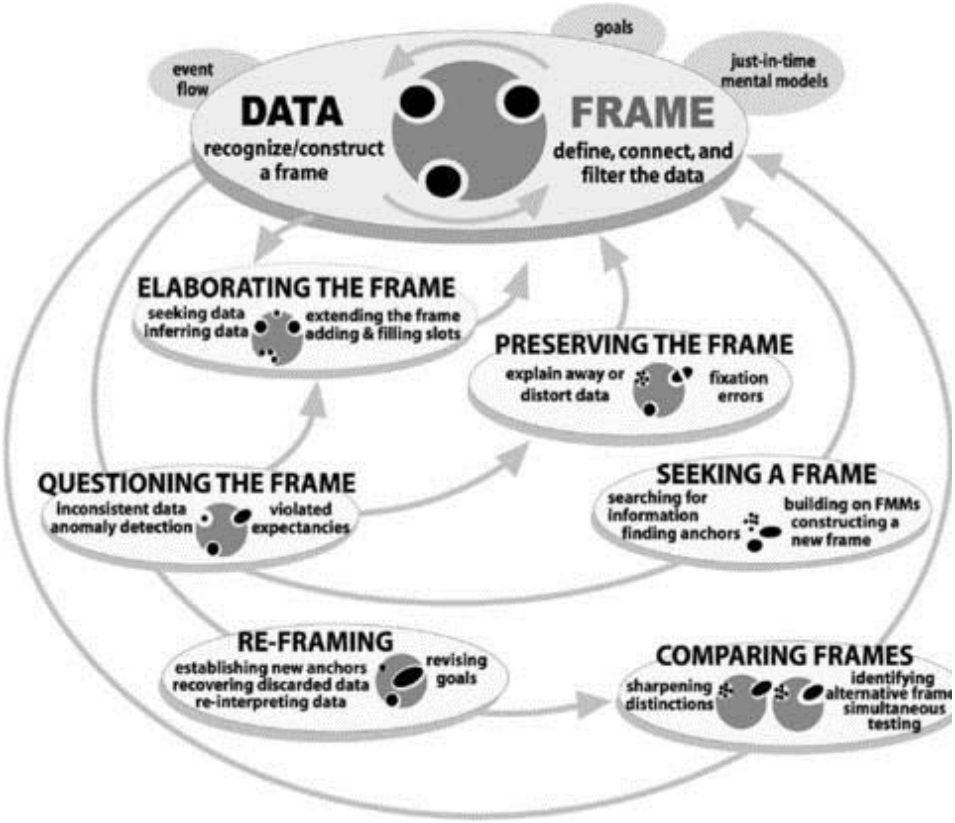

Figure 1. Data/Frame model of sensemaking (Klein, Phillips, Rall & Peluso, 2007).

Klein, Phillips, Rall & Peluso's (2007) DFM was based on a review of incidents arising from sensemaking studies of various functions including military navigation and operations planning, ICU nursing, firefighting and weather forecasting. In the current study we used the model as an *a priori* theoretical framework for classifying sensemaking actions. But given that we analysed a task in a different domain to those studied by Klein, Phillips, Rall & Peluso, we also allowed the analysis to extend beyond this model where appropriate in the manner of bottom-up thematic analysis (Braun & Clarke, 2006) for a complete analysis. Our intention was to leverage concepts within the DFM where appropriate, but to not be over-constrained by it as the exclusive conceptual lens.

USER-STUDIES

Two user studies provided data for training and testing the IFM. In both studies, participants were asked to identify leaders or 'influencers' in an academic field. These studies have been reported elsewhere (Kodagoda, Attfield, Wong, Rooney, & Choudhury, 2013; Pontis & Blandford, 2015). Here we describe them in the overview and discuss the analysis conducted for the current work.

Study A (using 'everyday' research tools)

Five post-graduate students studying Human-Computer Interaction (HCI) and five academics teaching and conducting research in HCI used a set of 'everyday' research tools to identify influential researchers in two academic domains. The tools included a Web browser, a word processor, and traditional writing tools (paper, pencil and post-it notes). Participants were given four tasks: identify (a) three current and (b) three future influential researchers in a HCI (a familiar domain) and the same in chemistry (an unfamiliar domain). Sessions were recorded using screen-capture software to document participants' interactions with the computer, and audio-recording devices to capture participants' thinking aloud (Ericsson & Simon, 1984) while performing the tasks. Sessions generally lasted a little over an hour. Then, semi-structured debriefing interviews were audio-recorded during which specific observations were discussed and associated thinking processes clarified (Charters, 2003).

Study B (using INVISQUE)

Six university librarians used a prototype tool called INVISQUE to identify influential researchers in information visualisation (Wong, Chen, Kodagoda, Rooney, & Xu, 2011). In INVISQUE, searches are submitted on an infinite canvas, and results presented as visual objects resembling index cards which appear in groups or 'clusters' (Figure 2). Both clusters and individual cards can be moved and rearranged to create new bespoke clusters. Within clusters, cards are laid in a format resembling points on a scatterplot with user-editable x and y axes. Search clusters are automatically named according to search terms and bespoke clusters are named by the user. For the study, INVISQUE was loaded with ACM SIGCHI conference papers from 1982 to 2011 (around 9000 publications).

After a briefing about the software and some familiarisation time with the tool (approximately 15 minutes), participants were asked to identify at least three influential authors in the field of information visualisation (unfamiliar domain). Sessions were recorded using screen-capture software to document participants' interactions with the computer, and audio-recording devices to capture participants' thinking aloud (Ericsson & Simon, 1984). While performing the tasks, automated interaction logs captured user activity and the researcher's notes were recorded. Sessions lasted around 40 minutes. Post-task interviews were audio-recorded and structured using SMART probes (Wong, Kodagoda, Rooney, Attfield, & Choudhury, 2013), an approach which uses retrospective cues based on the DFM.

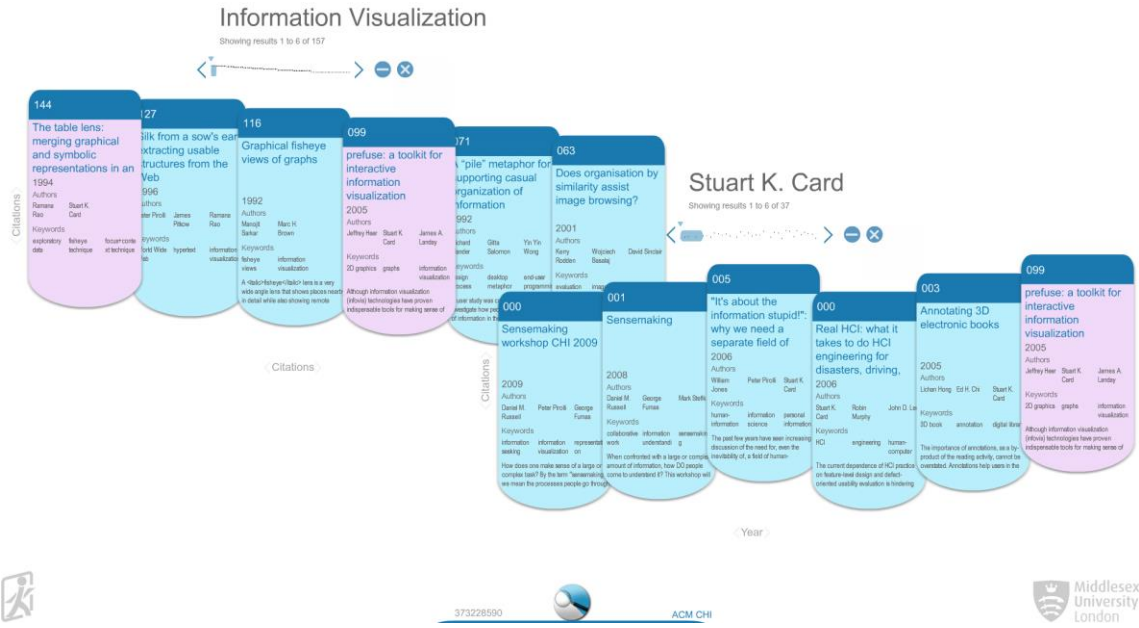

Figure 2. INVISQUE system showing two clusters. One cluster showing publications in relation to '***Information Visualization***' (left) and the other showing publications by '***Stuart K. Card***' (right).

Data Analysis and Coding

The two studies resulted in 480 minutes of recorded activity. From the 'unfamiliar' condition of study A, ten completed task runs were randomly selected for further analysis. From study B, data from all sessions were used. The recorded data protocols were analysed according to four dimensions:

Action: Provides a low-level trace of actions (analysis provenance) that participants performed to complete the task. Interaction logs for study A were manually reconstructed from the screen capture videos resulting in the identification of 23 unique low-level user actions, such as *scrolling*, *highlighting* and *opening (document)*. Interaction logs were automatically recorded during study B using INVISQUE's built in capability, resulting in the identification of 35 unique low-level user actions, such as *new_search*, *move_cluster_started*, *change_x_axis*.

Detail: Provides additional detail or qualification of a user action (the FMA only used names of possible influencers).

Indicator: Provides indicators that participants used as cues for *Frame Manipulation Actions*, (e.g. high citation, frequency of publication, and years of publication).

Frame Manipulation Action (FMA): Provides an associated sensemaking action inferred from (1) the low-level trace, (2) corresponding 'think-aloud' extracts, (3) retrospective interview data (*reasoning provenance*). The FMA was described using a coding dictionary (Table 2) which was shared and developed by researchers analysing both studies. Each code describes an action with a 'frame', where a frame is a user's internal representation of a situation. The scheme is based on the DFM as described in Klein, Phillips, Rall & Peluso, (2007), and reconciled against a slightly different account in Klein, Moon & Hoffman (2006). The coding scheme has the form of a shallow hierarchy.

Table 2

Coding scheme developed for coding data sets collected from studies A and B, based on Klein, Phillips, Rall & Peluso (2007) and Klein, Moon & Hoffman (2006).

Data and Frame	***Recognise data and construct a frame*** - Investigators develop their idea of what it means to be an influencer and/or the cues that might indicate one. ***Connect data with Frame*** - Investigators gain a moment of insight that may lead them to consider an individual as a potential influencer.
Elaborate a Frame	***Seek and Infer Data*** - Investigators look for information about someone they are currently considering as a candidate influencer. ***Fill Slots*** - Investigators discover new information which enhances their understanding of someone they consider as a candidate influencer, where the information is consistent with that hypothesis. ***Discard Data*** - Investigators discard information which they see as irrelevant to their hypothesis that someone is a candidate influencer. ***Extend Frame / add or refine slots*** - Investigators extend their concept of what it means to be or to identify an influencer. They develop the frame.
Question a Frame	***Detect Inconsistencies*** - Investigators discover information about someone they had been considering as a candidate influencer, which they see as violating expectations that the hypothesis had created. ***Track Anomalies*** - Investigators follow up on information which they see as inconsistent with a hypothesis they had been considering about someone being a candidate influencer. ***Judge Plausibility*** - Investigators reconsider the plausibility of a prior hypothesis that someone is a candidate influencer. ***Gauge Data Quality*** - Investigators find some information that they deem significant in terms of their hypothesis but they cannot fully trust the source. This may be because it conflicts with another more trusted source, there is missing data or there are known inaccuracies elsewhere in that source.
Compare Multiple Frames	***Comparing Frames*** - Investigators explicitly compare or explore more than one possible hypothesis related to an individual's status as a candidate influencer.
Stop Pursuing Frame Instantiation	Investigators stop exploring a given author.

The analysis was intentionally limited to frames that corresponded to participants' interpretation of someone as an 'influencer'. Two researchers analysed data from the two studies, one analysing study A and one analysing study B. To enhance reliability, the dictionary included definitions of each code contextualised to the type of user-task under study. Reliability was checked by each researcher blind coding data from a randomly selected participant from the other study. In total, 100 instances were coded by both researchers and inter-rater reliability statistics calculated on the codes using Cohen's kappa (Cohen, 1960) showed a reliability coefficient of $\kappa = 0.837$. Kappa values above .80 are generally considered good (Brett & Jeanne, 1998). This comparison also stimulated discussion about the DFM and resulted in some development of the coding scheme. Comparing coded data, and having meetings during the data analysis process supported coding agreement.

Table 3 provides an illustrative extract of the analysis, using an example from study B. During the early stages of the extract the participant develops, or gives expression to, a notion of what it means to be an influencer in the given field. This includes a search for 'information', a search for 'visualisation', combining these using Boolean AND, and ordering a document cluster by citation count. These actions were coded as *Recognise data and construct frame*. The participant also discards the old clusters, coded as *Discard data*.

Table 3

Example of interaction log used in the study.

Action	Detail	Indicator	FMA
New_Search	keyword: information, no of results: 2112		Recognise data and construct a frame
New_Search	keyword: visualization, no of results: 556		Recognise data and construct a frame
Boolean_AND	cluster 1: visualization, cluster 2: information		Recognise data and construct a frame
Close_Cluster	cluster: visualization		Discard data
Close_Cluster	cluster: information		Discard data
Change_X_Axis	selected: citations, x: citations, order: descending	[High citation]	Recognise data and construct a frame
Keyword_Highlight	cluster: information visualization, node: 191776, keyword: <author name>	Author: <author name> [High citation]	Connecting Data with Frame
Keyword_Highlight	cluster: information visualization, node: 191776, keyword: <author name>	Author: <author name> [High Citation, number of publications over time]	Detect inconsistencies

The participant then selects the author field of the most highly cited paper with the effect of highlighting all publications by that author. Since the participant is considering a specific author as an influencer this is coded as *Connecting data with frame* and the *Indicator* is shown as *<author name>*, *high citation*. Looking at the scatter plot the

participant then observes that there are only a few publications by this author. The *FMA* is coded as *Detect inconsistencies* and the *Indicators* are *<author name>, high citation, number of publications over time*.

TRAINING THE INTERACTION FRAME MAPPER

We used results of this analysis to train the IFM. The training aimed to inductively identify relationships between interaction sequences (*analytic provenance*) and associated reasoning processes (*reasoning provenance*). The former was represented by the low-level user interaction events captured by the *Action* feature. The latter was represented by interpretive coding captured by the FMA feature.

Feature characteristics

Given differences in the tools used by participants in each study and the different action sets these generated, training and testing were conducted for each study independently. This helped abstract away the level of detail classifying cognitive actions into the types defined by the DFM. However, a common feature set was used for both.

Features

The IFM used four features and one class label. We divide these into two categories: *general* and *task-based*. General features are not customised to a particular task. *Action* (an interaction with the computer made by the participant) is the only general feature used. As with all features, an interaction should be recordable by the computer, so that the process can in principle be automated. For this study, the interaction logs for study B were captured automatically because the interface permitted this operation. However, study A, being a freeform web exercise, did not record interaction details; hence, these were annotated afterwards by a human coder.

Task-based features aimed to improve the reasoning recovery by adding additional context. The task in question for both studies was to determine influential individuals. From the manually labelled protocol we extracted a subset of features that could reveal users' reasoning for this task. As the intention is to provide a method that can create these features automatically during software usage, they must not require human observation or inference to be populated. Pre-processing of data included converting features to numeric representations. *Detail Person* and *Person Previous* were simple binary features; the *Interaction* and *Indicator* features were mapped from a set of categorical labels to a set of integers. This representation was fine for scale-invariant models such as the Random Forest (RF) (Breiman, 2001) and Hidden Markov Model (HMM) (Rabiner & Juang, 1986), however the Support Vector Machines (SVM) (Cortes & Vapnik, 1995) algorithm is not scale invariant so the integers were standardized to have mean 0 and a variance of 1. The task-based features are described below (Table 4):

- **Detail Person:** if the additional information field contains a person entity (binary) determined by named entity extraction;
- **Person Previous:** if the field contains a person reference where they are mentioned previously (binary);

- **Indicator:** an explicit use of a specific indicator or cue (categorical). Indicators here mean a property that the participant is using to determine if someone is influential. The indicator list was determined before the study. This data was reconstructed for study A and available from interaction logs for study B;
- **Example influence indicators:** citations, publications, both citations and publications, awards or discoveries;

Table 4

Sample excerpt from study B training data.

Interaction	Detail person	Person previous	Indicator	fma*
Change_X_Axis	False	False	Missing	2
Bookmark_Added	False	False	Citationpublication	2
Keyword_Highlight	True	False	Citationpublication	2
Keyword_Highlight	True	False	Citationpublication	2
Keyword_Highlight	False	False	Citationpublication	2
Keyword_Highlight	True	False	Citationpublication	2
Delete_Node	False	False	Missing	4

***class label**

Data

Once rows without *Action* or *FMA* data were removed, the training set consisted of 403 rows from study A and 690 from study B. Due to the difference in tools used in the different studies and the resulting action sets, we trained different classifiers for each study. According to Hastie, Tibshirani & Friedman (2001), it is difficult to give a general rule of thumb on how much training data is enough, however a commonly presented rule of thumb is ten times as many training instances as tuneable parameters in the model. Both studies have at least ten times as many training instances as features so although the sample sizes are not large they should remain generalizable.

Classification Technique

We used three different machine learning classifiers to test the hypothesis that computer-based interactions can provide information to aid in recovering reasoning. We tested against a control classifier which naively predicts the most frequent class in the data. Referred to as the *no information classifier*, it represents a baseline that any model should aim to beat.

SVM (Cortes & Vapnik, 1995) and RF (Breiman, 2001) classifiers were used since these have exhibited high general performance for classification tasks. However, neither consider feature sequence which may be important, and so we also used a HMM (Rabiner & Juang, 1986) classifier. We used a 2-fold cross validation approach to evaluate classifier performance and to configure the chosen models correctly.

Ideally, a held-out set would be preserved but we used all the data that could be used, because of the difficulty in obtaining more data (due to the time-intensive manual coding of protocols). For the classification task, we used a small number of training examples and small feature set; hence we expected classifiers that perform well with these restrictions to perform better.

The SVM used a linear kernel with the cost parameter c set to 1—an iterative stepped cross-validation found no improvement for larger values. Given the small feature and sample space, a Gaussian kernel such as Radial Basis Functions (RBF) might be expected to perform better. However the linear kernel performed best in cross-validation, suggesting the data is linearly separable.

RF classifiers tend to perform worse with small data sets and where there is little variation within instances. This can be due to the resampling method not being able to create different decision trees because of high bias. Whilst our sample size was small, there was a reasonable variance, particularly for study B.

As discussed above, sequence may have a significant role in recovering reasoning. It is worth noting that the HMM confers a limited view of sequencing due to the Markovian independence assumption. This may be stated as: the future is independent of the past given the present. On this view of sequencing, the information needed to determine the next step is the state you are currently in (computed via probability of emitting an observation) and what the likelihood of transition from that states is. This is a simplification that has worked well in other areas, including speech recognition and signal processing. It is possible that the reasoning activity sequencing problem is more complex than this simplifying assumption, but evaluating this classifier should at least determine if this simple view of sequencing confers any useful information for this scenario. The HMM for each cross-validation fold was trained using a supervised learning approach. Classification was performed on a vector of observed interactions. The most likely states are calculated from these observations using the Viterbi Path (Viterbi, 1967).

EVALUATION RESULTS

We evaluated the performance of our approach by looking at the overall classification accuracy of each of the three models and both studies. Both studies exhibited a skewed class distribution; hence, accuracy is a misleading evaluation metric. Instead we use the weighted average F1 measure—the F1 measure per class weighted by the number of class instances. F1 score is the harmonic mean of precision and recall and is only high where both precision and recall are high. The weighted average F1 can conceal classes in which the models/studies performed well or not so well. Hence we also examine the per-class breakdown in Class-based F1 measure analysis. We present a comparable study in the section *Class-based F1 Measure Analysis*.

Overall accuracy of both studies

Both studies created a skewed class distribution for FMA (Figures 3.1 & 3.2). There are also some differences. Since inter-rater assessment was performed these differences are likely to stem from the nature of the tasks being performed. This leads to higher instances of the *Fill Slots* class and lower counts of the *Judge Plausibility* class.

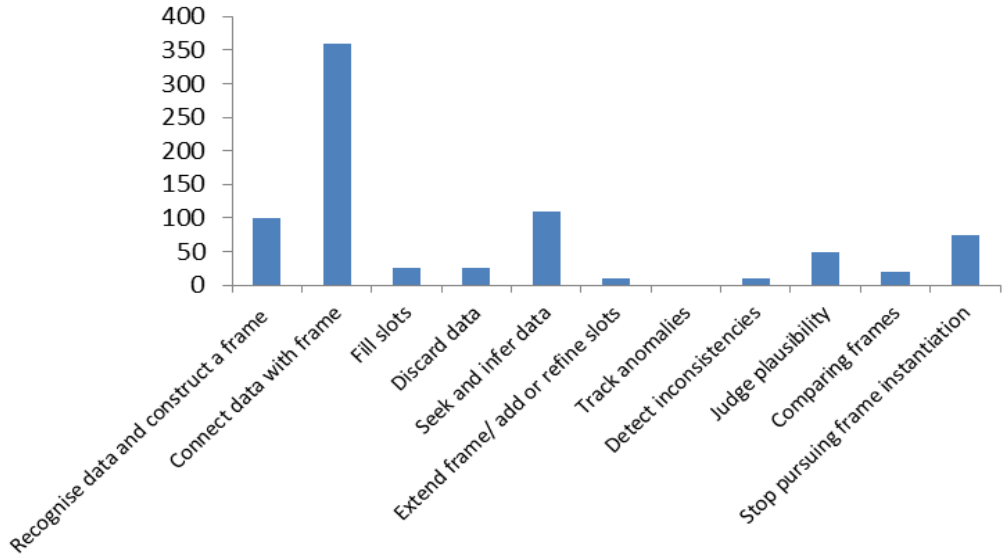

Figure 3.1. FMA Class Distribution for study A.

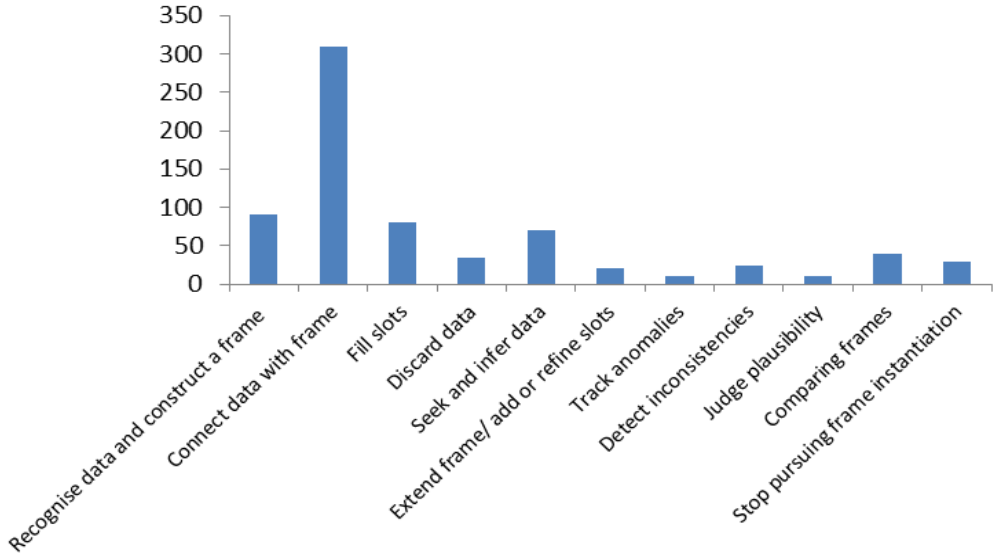

Figure 3.2. FMA Class Distribution for Study B.

The key comparison for each of the three approaches is whether they significantly outperform the *no information* classifier. Statistical significance is determined using a one-sided binomial test on the weighted F1 score with α at 0.05. The results for study A and B are shown in Table 5, and Figures 4 and 5. None of the candidate models in study

A confer a benefit over the no-information rate. The SVM model is close to the performance of the most frequent classifier as it mostly predicts the most frequent class.

The distinction between experts and novices, and domain familiarity was explored statistically as a precautionary step for study A. There were no significant differences in the observed sample set and so, for the purposes of the study, they could be treated as derived from the same population.

Table 5

Study A and B, weighted average F1 measure model comparison.

Study	SVM	Random Forest	HMM	No Information
A	0.4817	0.0258	0.0744	0.4717
B	0.4602	0.4043	0.2510	0.2928

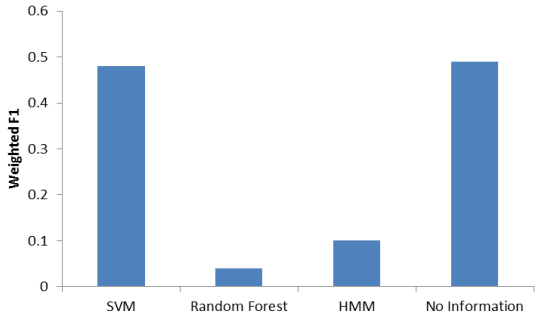

Figure 4. Study A, weighted average F1 measure model comparison. The Support Vector is the best model, however none of the models significantly outperform the no-information-rate classifier ($p<0.05$). The Random Forest performs particularly badly because of the high bias present in the data. As will also be seen in study B, the HMM does not perform well on this task.

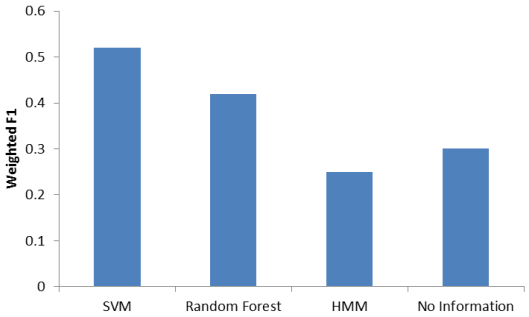

Figure 5. Study B, weighted average F1 measure model comparison. The Support Vector is the best model and it does significantly outperform the no-information rate classifier ($p < 2.2e\text{-}16$). The random forest performs better here as there is less bias in the main feature set. It also is significantly better than the no-information rate ($p=2.272e\text{-}12$) The HMM does better here but not as well as the no-information classifier.

Some classes seem to be more predictable than others, and this is particularly evident with study B. Study A is able to predict the most frequent class well because it has the best mapping of interaction (*scrolling* in this case) to the *Connect Data with Frame* class. This is not true for other classes in this model; most have equal amounts of each interaction feature observed in the class training data. There is hence little for the classifiers to differentiate the other classes. The most frequent interaction feature (*scrolling*) is spread uniformly across the class features and so provides little information to improve FMA prediction.

The Random Forest does not perform well. The data exhibits a high bias and the effect of bagging the data into different decision trees cannot be expected to work well. Surprisingly the HMM also performs quite poorly. This suggests that at least this simplification of sequencing does little to aid in recovering reasoning processes. The HMM is trained off a combined training set from different users due to the nature of our

dataset. Each user performed one experiment. Having multiple experiments for the same user may improve this method by training over an individual's reasoning patterns.

Study B performs better on this classification task. Both the SVM and RF approaches significantly outperform the no-information classifier. Results and p-values are shown in Figure 5. SVM performs better than RF (as in A), which is what we would expect given the sample and feature space. The RF method performs significantly better mainly due to increased variation in the feature space. The distributions for each study's main feature *action* are shown in Figures 6 and 7 and Tables 6 and 7 respectively. The HMM again performs poorly by comparison, and again this suggests that a stronger mapping exists between certain actions and reasoning activities than in the sequencing of reasoning activities.

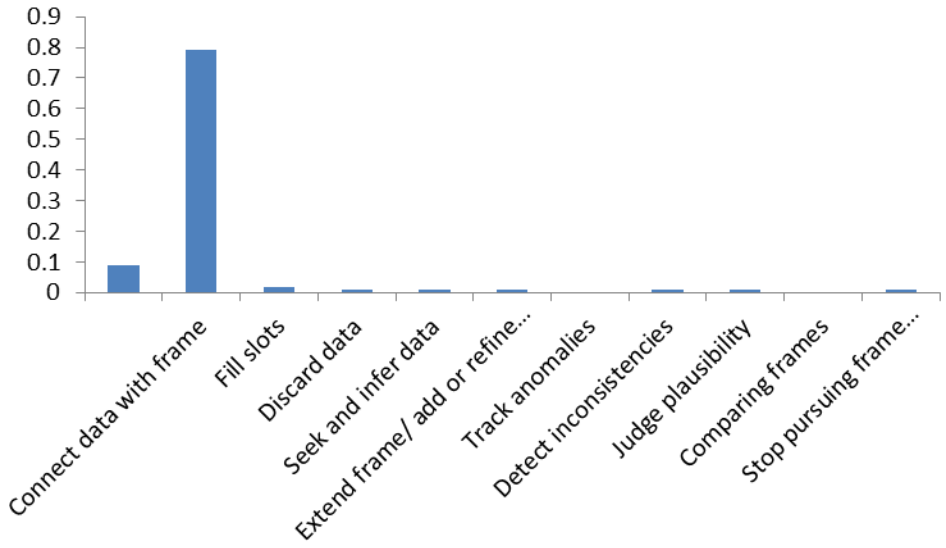

Figure 6. Study A, SVM per-class classification performance. Only results for SVM are shown as was best performing model. Study A is no better than the no-information classifier. It exhibits high bias due mostly to little differentiation in the action set in comparison to study B.

Table 6

Most frequent interactions observed for classes in study A, sorted by count. These classes are the best performing from study A.

FMA	Interaction	N	Interaction	N
Recog.	Scrolling	30	Clicked_On	13
Conn.	Scrolling	206	Clicked_On	98
Discard	Scrolling	12	Going_Back_To	7
Seek	Scrolling	54	Clicked_On	28
Stop	Going_Back_To	12	Moving_Mouse	8

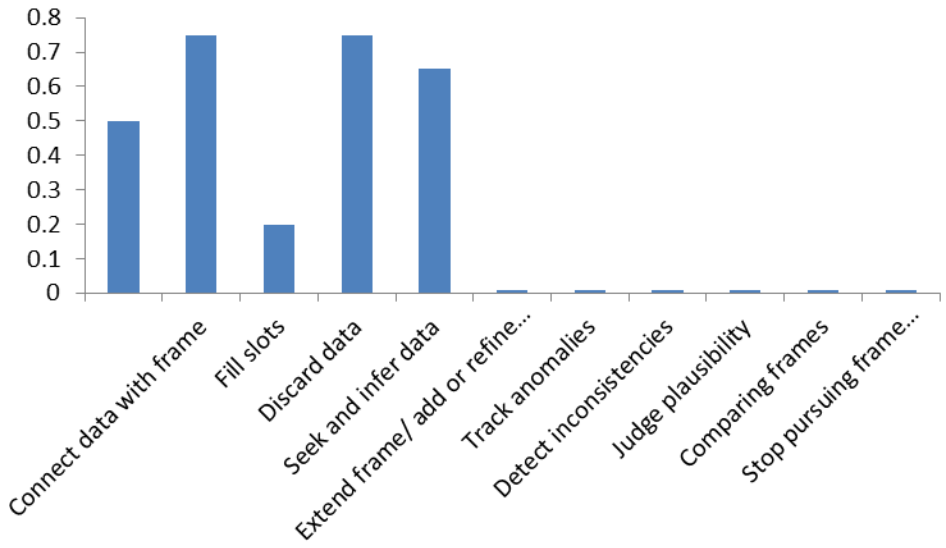

Figure 7. Study B, SVM per-class classification performance. Study B outperforms the no-information classifier in general and has reasonable performance for five classes. Study B actions tie more closely to the FMA and in the case where those actions have a distinct mapping to a FMA the classifier can predict the activities.

Table 7

Most frequent interactions in study B, sorted by count. These classes are the best performing from study B.

FMA	Interaction	N	Interaction	N
Recog.	Change_X_Axis	25	Move_Cluster	23
Conn.	Move_Cluster	97	Keyword_Highlight	93
Discard	Close_Cluster	12	Cluster_Removed	9
Seek	New_Search	49	Scrolling	14
Stop	Close_Cluster	15	Going_Back_To	5

The overall weighted F1 score is not very high for any method, even SVM. However this is a difficult task and a significant improvement over the baseline highlights the potential. The weighted F1 scores hide the individual class performance. In the following, we examine per-class F1 scores.

1.1 Class-based F1 Measure Analysis

The SVM performed best out of the candidate classifiers. For brevity, we only examine the per-class F1 scores for the SVM classifier.

Figure 6 shows the SVM multi-class F1 performance for study A. SVM performs poorly on all classes except for class 2 (*Connect Data with a Frame*). The F1 for class 2 is moderately high, but this is due to it being the most frequent observation. This classifier exhibits high bias and predicts class 2 in the majority of cases, and provides no gain over the no-information rate. Study B shows reasonable F1 performance in five classes: *Recognise Data and Construct a Frame, Connect Data with a Frame, Discard Data, Seek*

and Infer Data and *Stop Pursuing Frame Instantiation.* Examining the observed actions for these five classes in study A shows that it would be difficult for any classifier to produce a good result, due to low differentiation between classes (Table 6). These indicate the difficulty for a classifier to predict reasoning activity from the interactions that do not have an underlying mapping to the sensemaking process.

Figure 7 shows the multi-class F1 performance for study B data. Five of the classes display reasonable performance: *Recognise Data and Construct a Frame, Connect Data with a Frame, Discard Data, Seek and Infer Data* and *Stop Pursuing Frame Instantiation*. Examining the top interactions for these classes (Table 7) we see that the top interactions are distinct, aiding classifier performance. We deduce from this that when the interface is constructed with sensemaking in mind it is easier to reconstruct at least part of the reasoning. We could speculate the intent of these interactions; for example with *Recognise Data and Construct a Frame* we observe users ordering data by some dimension in x axis. This is as a way of exploring the data.

Poor performance in other classes may be because they fail to map directly to interface actions or we lack the data/features to predict them. It is worth noting that INVISQUE does not provide a method for tracking anomalies, marking inconsistencies or annotating plausibility. The poor results for these classes may be because the software does not clearly map these activities to actions on the interface. One of the results was surprising. INVISQUE provides a method of comparing different clusters at once. We would have expected this to produce better results for the *Comparing Multiple Frames* class. However it is not apparent in the interaction logs when a user is manipulating multiple clusters. Future work will investigate changing the interaction logging to reflect this.

Figures 8 and 9 show the action frequency distributions for both studies. Although these have similar shape, the top action for study A, *no_explicit_interaction* is uniformly distributed among all of the FMAs in the training data and hence provides no information to the classifiers.

Dou et al., (2009) provide the only directly comparable work to this. They employed a manual analysis with multiple human coders to recover findings, strategies and methods of financial analysts from interaction log data. Their strategies and methods are similar to what our study is trying to determine and hence we use those values for comparison. An additional caveat is that we have used weighted F1 in our example, not accuracy. Also we do not know if there is class skew in this comparison example. The accuracy of the SVM model on study B data is 59.4%. A naive comparison shows that the INVISQUE based classification accuracy of 60% is reasonably close to the levels produced by human experts in completing a similar task to our automated approach (Figure 10).

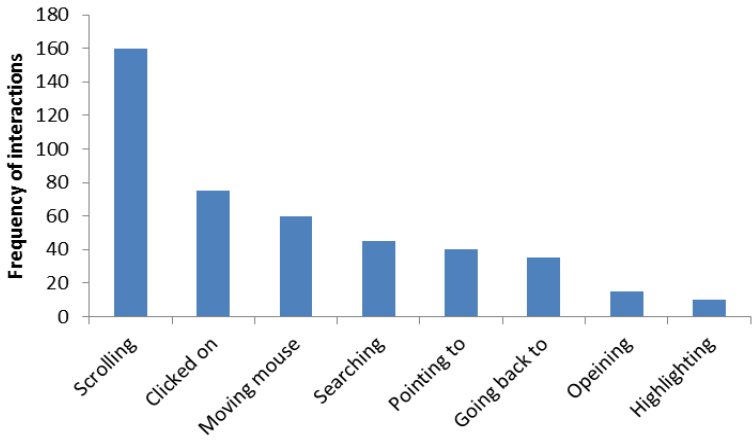

Figure 8. Study A, interaction frequency distribution. The interaction set for this study contains 23 distinct actions; only those with more than 5 occurrences in the data have been presented for presentation reasons.

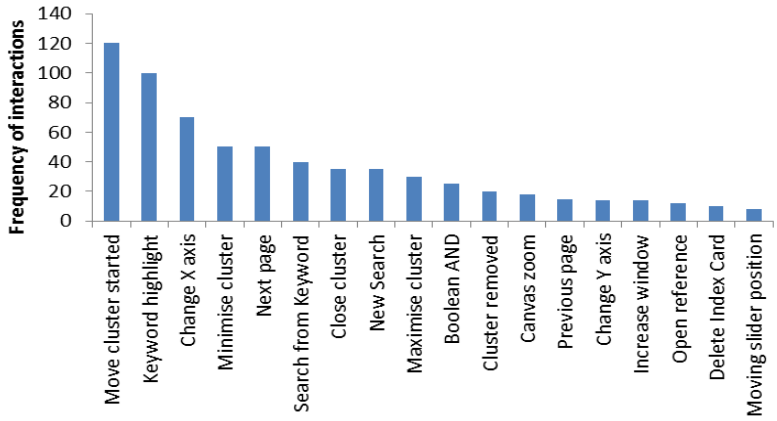

Figure 9. Study B, interaction frequency distribution. The interaction set for this study contains 35 actions; these are filtered as in *Figure 6*.

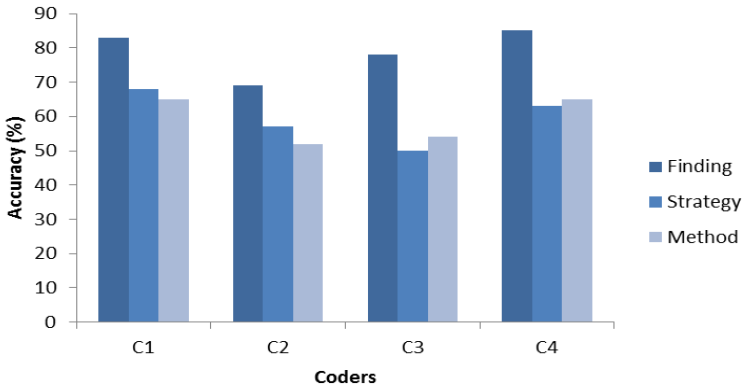

Figure 10. Comparison based classification accuracy. Image used from Dou et al., (2009). Details accuracy of multiple coders in determining reasoning from interactions. C[1-4] are different human coders.

DISCUSSION AND FUTURE WORK

We present a novel method for reconstructing *reasoning provenance* from *analysis provenance* records using a sensemaking model to map from actions to reasoning. We used a modified version of Klein, Phillips, Rall & Peluso's (2007) DFM to represent sensemaking or reasoning activities. Multiple classifiers were used to generate *reasoning provenance* artefacts from interaction logs and task-based features such as frame cues or 'anchors'. We used protocols from two studies based around identifying influential individuals to evaluate the approach. The coding was part based on participants' think aloud while performing the task, which necessarily involved details of the content of frames, or the semantics of the domain. However, we abstracted away from that level of detail in order to classify cognitive actions into the types defined by the DFM. This kind of abstraction is important for an approach which could be applied to multiple domains.

After we trained the mapper with data obtained from the two user studies, we found that the SVM was the best model from cross-validation of labelled data. Sequences of actions were evaluated using a HMM. This was found to be less effective from mapping interaction to reasoning activity, at least with our simplification of sequencing.

Study A did not provide a reliable mapping from interaction to sensemaking activity with its classifier exhibiting a high bias towards omitting the most frequent class. Study B did show a reliable mapping between interface actions and sensemaking activity with the weighted F1 measure significantly above the *no information* rate. The study B classifier also did reasonably well across classes with some exceptions. To understand the exact cause of this effect would require additional data in which adequate class data was available for all classes under test. However we did find that a class with a small number of instances could predict well, potentially indicating that the action (& context) mapping to reasoning process is not uniform across different types of reasoning activity. However, we may be able to explain this difference in terms of the differences in the tools used for the two studies. In study A, participants used a Web browser, word processor, and traditional writing tools (paper, pencil and post-it notes). Actions revolved mainly around the use of the Web browser, and hence observed actions predominantly reflected information seeking activity. Information seeking is a reciprocal partner of sensemaking, but may be too far removed from sensemaking related cognition (*i.e.* internally theorising) to support reliable prediction. In study B participants used INVISQUE which allowed them to visually organise information objects (documents in this case) in ways which might more directly reflect underlying cognitive organisation. For example, they organised documents into clusters according to the various authors (frames) that they were considering as candidate influencers. More detail is provided in the original reporting of study B in Kodagoda, Wong, Rooney & Khan (2012), Kodagoda, Attfield, Wong, Rooney, & Choudhury (2013). Similar observations were also made in Rooney, Attfield, Wong, & Choudhury (2014).

The value of study A was in the contrast it provides to study B—which was significant. We used both studies together to explore conditions under which interaction events might predict FMAs, and we found this result in one study and not in the other. This helps us in launching a post-hoc explanation for why, which we would not have without study A.

And given that study B gave a significant result, we argue that the findings are not inconclusive. With the two studies, we are able to offer the explanation that the free form canvas design used in study B allowed users to organise information in groupings which acted as proxies for internal cognitive constructs such as theories or hunches. This externalisation enables a form of distributed cognition, or 'distributed sensemaking' which we suggest improved performance on study B. This performance is not at a level which is usable in practice, but it provides a starting point and useful direction for further research.

The implication is that tools will be more successful at predicting *reasoning provenance* where there is a closer mapping between interface actions and cognitive organisation. Arguably, such tools are better suited to sensemaking tasks. This leads to a further conclusion: if the ability to reconstruct *reasoning provenance* from observable interaction is treated as a test of the quality of the mapping between observable actions and sensemaking cognition, then this might also be used as the basis for a tool evaluation method. If a good tool is one in which cognitive actions and organisation map closely to interface actions and organisation, then this should be indicated through the ability to predict one from the other.

There are some notable limitations to our approach. Due to the effort required to recreate this data from previously coded protocols, our sample size is small. The main bottleneck here is the time spent by the coder in reviewing audio, visual and textual protocols to infer the reasoning activity being performed at any given time. The limited amount of data also means that some classes are under-represented in the data with very low instance counts, some less than three. This makes it difficult for the classifier to distinguish between instances. The creation of more training data would be a requirement in furthering our model. An additional limitation is that a classifier must be trained per action set (user interface). A solution to this is non-trivial; however, a work-around would be to only use a custom system designed for this purpose, for example, a visual analytics toolkit. This would allow a single interaction set.

To develop this work further we plan to create an extension for the INVISQUE analysis tool that allows our data to be captured automatically and which could integrate a trained task classifier into the system. This would allow a more robust evaluation of the technique where FMAs are predicted by the system and a post-session survey of the participant could check the model accuracy. We also plan to add features into the model. These could be based on additional sensing capabilities such as eye tracking for attention discovery or textual analysis of notes entered during an analysis. Temporal effects such as the duration of an action will be explored.

The advantage of being able to predict FMAs is that these could form the basis of a model of *reasoning provenance*. This could then be used as the basis of visual representations that could support user-reflection on the process (*reflection-in-action*). It could also form the basis of representations to inform third party audit and training etc. Producing such models and representations is out of scope of the current paper, but the work presented in the paper is a necessary step in producing such models.

This paper contributes to insights about (i) design and development of interfaces to enhance the external proxies of cognition/sensemaking, (ii) a way of capturing *analytical provenance* with richer context, (iii) training a FMA to infer *reasoning provenance* from captured *analytical provenance* using machine learning, and (iv) using a sensemaking model as a framework.

The contribution is not only the use of machine learning to infer users' *reasoning provenance*, but also a contribution to the design and development of interactive visualisation systems and visual analytics tools aimed at facilitating the recording of *analytic* and *reasoning provenance* based on the DFM of sensemaking. A potential additional feature would be to present the current *reasoning provenance* graph to the user and allow them to view their reasoning path, correct and refine it, which could update the trained model.

ACKNOWLEDGEMENTS

The authors would like to thank all participants from both studies A and B who agreed to be part of the experiments. The research reported has been supported through the UK Visual Analytics Consortium (UKVAC) that has been funded by Her Majesty's Government and the U.S. Department of Homeland Security, by project 4112-46065.

REFERENCES

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi:10.1191/1478088706qp063oa

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Brett, Jeanne, D. L. S. and A. L. L. (1998). Breaking the Bonds of Reciprocity in Negotiation. *The Academy of Management Journal*, *41*(4), 410–424.

Charters, E. (2003, July 1). The Use of Think-aloud Methods in Qualitative Research An Introduction to Think-aloud Methods. *Brock Education Journal*.

Chen, Y. V., Qian, Z. C., Woodbury, R., Dill, J., & Shaw, C. D. (2014). Employing a Parametric Model for Analytic Provenance. *ACM Transactions on Interactive Intelligent Systems (TiiS) - Special Issue on Interactive Computational Visual Analytics*, *4*(1). doi:10.1145/2591510

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. doi:10.1177/001316446002000104

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297. doi:10.1023/A:1022627411411

Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R., & Chang, R. (2009). Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications - Special Issue on Sketching Tangible Interfaces Augmented Reality on Mobile Phone*, *29*(3), 52–61. doi:10.1109/MCG.2009.49

Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data*. MIT Press.

Gotz, D., & Zhou, M. X. (2008a). *An Empirical Study of User Interaction Behavior during Visual Analysis*.

Gotz, D., & Zhou, M. X. (2008b). Characterizing users' visual analytic activity for insight provenance. *2008 IEEE Symposium on Visual Analytics Science and Technology*, 123–130. doi:10.1109/VAST.2008.4677365

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media. doi:10.1007/978-0-387-84858-7

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems*, *21*(5), 88–92. doi:10.1109/MIS.2006.100

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). Expertise Out of Context: A Data-Frame Theory of Sensemaking. In R. R. Hoffman (Ed.), *Expertise out of context: Proceedings of the Sixth International Conference on Naturalistic Decision Making* (pp. 113–155). Taylor & Francis Group. doi:10.4324/9780203810088

Kodagoda, N., Attfield, S., Wong, B. L. W., Rooney, C., & Choudhury, S. T. (2013). Using interactive visual reasoning to support sense-making: implications for design. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 2217–26. doi:10.1109/TVCG.2013.211

Kodagoda, N., Wong, B. L. W., Rooney, C., & Khan, N. (2012). Interactive visualization for low literacy users: from lessons learnt to design. In *CHI '12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1159–1168). doi:10.1145/2208516.2208565

Nardi, B. A. (1995). *Context and consciousness: activity theory and human-computer interaction*. (Bonnie A. Nardi, Ed.). Massachusetts Institute of Technology.

North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B., & Fink, G. (2011). Analytic provenance. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (p. 33). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979570

Pontis, S., & Blandford, A. (2015). Understanding "influence:" an exploratory study of academics' processes of knowledge construction through iterative and interactive information seeking. *Journal of the Association for Information Science and Technology*, *66*(8), 1576–1593. doi:10.1002/asi.23277

Rabiner, L. R., & Juang, B. H. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16.

Roberts, J. C., Keim, D., Hanratty, T., Rowlingson, R. R., Walker, R., Hall, M., … Varga, M. (2014). From Ill-Defined Problems to Informed Decisions. In *EuroVis Workshop on Visual Analytics*.

Rooney, C., Attfield, S., Wong, B. L. W., & Choudhury, S. (2014). INVISQUE as a Tool for Intelligence Analysis: The Construction of Explanatory Narratives. *International Journal of Human-Computer Interaction*, *30*(9), 703–717. doi:10.1080/10447318.2014.905422

Schon, D. A. (1983). *The Reflective Practitioner: How Professionals Think In Action*. Basic Books.

Starbuck, W. H., & Milliken, F. J. (1988). *Executives' perceptual filters: What they notice and how they make sense*. (Donald Hambrick, Ed.). Greenwich: The Executive Effect: Concepts and Methods for Studying Top Managers CT: JAI Press.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*(2), 260–269. doi:10.1109/TIT.1967.1054010

Weick, K. E. (1995). Sensemaking in Organizations. *Organization Studies*, *3*, 248.

Wong, B. L. W., Kodagoda, N., Rooney, C., Attfield, S., & Choudhury, S. (Tinni). (2013). Trialling the SMART approach: identifying and assessing sense-making. In *Proceedings of the Human Factors and Ergonomics Society 57rd Annual Meeting*. San Diego, California, USA: Human Factors and Ergonomics Society. doi:10.1177/1541931213571048

Wong, W., Chen, R., Kodagoda, N., Rooney, C., & Xu, K. (2011). INVISQUE: Intuitive Information Exploration through Interactive Visualization. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11* (pp. 311–314). New York, New York, USA: ACM Press. doi:10.1145/1979742.1979720