INTERPRETABLE MACHINE LEARNING FOR PRIVACY-PRESERVING IOT AND PERVASIVE SYSTEMS

Benjamin Baron, University College London
Mirco Musolesi, University College London and The Alan Turing Institute

Abstract — The presence of pervasive computing in our everyday lives and emergence of the Internet of Things, such as the interaction of users with connected devices like smartphones or home appliances generate increasing amounts of traces that reflect users' behavior. A plethora of machine learning techniques enable service providers to process these traces to extract latent information about the users. While most of the existing projects have focused on the accuracy of these techniques, little work has been done on the interpretation of the inference and identification algorithms based on them.

In this paper, we propose a machine learning interpretability framework for inference algorithms based on data collected through IoT and pervasive systems and we outline the open challenges in this research area. Our interpretability framework enable users to understand how the traces they generate could expose their privacy, while allowing for usable and personalized services at the same time.

Keywords: Interpretable Machine Learning, Internet of Things, Pervasive Computing, Privacy, Personal Data, Location Information.

1 | *Introduction*

With the emergence of connected devices (*e.g.*, smartphones and smartmeters), Internet of Things (IoT) and pervasive systems generate growing amounts of digital traces as users undergo their everyday activities. These traces are crucial to service providers to understand their customers, to increase the degree of personalization, and enhance the quality of their services. For instance, presence traces stemming from Oyster-like smartcards help transportation providers understand the commuting patterns of users; the usage statistics of home appliances can be used to improve energy efficiency; on-street cameras provide police officers with new ways of investigating crimes; content generated through mobile and wearables (such as posts in online social media or GPS running routes in specialized websites such as those for fitness) can be used to provide tailored content to individuals; bank transaction logs can be used to spot unusual activity in accounts.

However, sharing these digital traces generated by pervasive computing and IoT devices with service providers might raise concerns with regards to privacy. Indeed, the processing and analysis of these digital traces can surface latent information about the behavior of the users. While service providers have to store the user-generated data in large databases that guarantee a certain level of privacy (*e.g.*, storing the traces in an anonymized manner using pseudonyms instead of the real user's name and surname), third parties such as advertisers that have access to the traces can leverage machine learning techniques to reveal personal information about the users and expose their privacy. This includes inferring personal information about users and identifying a single individual from a collection of user-generated traces. Moreover, presence traces might reveal information about the significant places routinely visited by the user, enabling the service provider to infer a wide range of personal information, including the user's place of residence and work and their future locations. Presence traces can also be used to identify a specific individual in a population.

The focus of the existing work has been on the algorithmic and computational performance of the techniques to infer personal information and identify users from their digital traces (see for example [1]). Little interest has been shown on the interpretability of the results and models associated with these specific tasks, and the implications for the privacy exposure of the users, given the nature of the inference itself. In this context, interpretability of machine learning algorithms consist in giving effective and intelligible explanations of the identification task and model to the user [2]. As a result, the explanations must provide an intelligible representation to the users about what the model knows, how it knows it, and what it is doing about it. The need for the interpretability through effective explanation of the learning and inference process leading to certain set of outputs is twofold: (*i*) it will help users understand why their privacy has been exposed, and (*ii*) it enables users to trust the model's predictions, in order, for instance to follow the recommendations to take the necessary actions to protect their privacy in the future.

In this paper, we discuss the challenges related to the design of a generic interpretability framework with the goal of supporting interpretation of machine learning techniques that are adopted to expose the privacy of individuals through personal data inference and user identification. Our contributions are threefold. We state the interpretability and privacy requirements of an effective interpretability framework for privacy preserving IoT and pervasive systems before detailing the functionalities of its components, with a focus on feature selection methods as they are crucial when it comes to present the explanations to the users. We present a case study where we detail a prototype of the interpretability framework that relies on machine learning classifiers with the goal of identifying users from samples of their presence traces. In particular, by considering presence traces of users from Foursquare, a location-based social network, which given its availability and scale, enables the reproducibility of the present study. Finally, we present the open challenges in this area, discussing a potential research agenda for collaboration across the pervasive computing, human-computer interaction and machine learning communities.

2 | *Towards a Privacy-Oriented Intepretability Framework for IoT and Pervasive Systems*

2.1. Privacy Requirements for Digital Trace Inference Interpretability

The main goal of the proposed interpretability framework is to automatically generate explanations of inference tasks on personal data. The explanations should be helpful to the end-users so that they have an understanding on how the underlying machine learning model works and behaves depending on the inputs (*i.e.*, the data provided by the end-user). User understanding is fundamental, as it improves the trust and acceptance of predictions and recommendations given by the model. In particular, this has been shown to be effective in domains such as decision making and recommender systems [3]. In this section, we present the different requirements concerning the design of a privacy-oriented intepretability framework for machine learning techniques applied to digital traces analysis and inference. Several works have conducted extensive user studies in order to determine the main requirements of explanations about machine learning models and their predictions given to users [2, 4, 5].

This is usually studied from a usability perspective [6]. In particular, researchers and practitioners focus on the so-called *gulf of evaluation*, *i.e.*, the gap between representations that can be directly perceived and interpreted by a user provided by a system and her/his expectations and intentions. Another key aspect considered in usability studies is the *gulf of execution*, *i.e.*, the gap between a user's goal related to a specific action and the means to execute that goal. While the main requirements focus on addressing both the gulf of evaluation and the gulf of execution, we argue the case for an additional privacy requirement. We support the choice of the requirements by providing examples through a discussion of a case study that we will present later in Section 3. The scenario of the case study describes a machine learning interpretability framework that provides interpretations for inference of personal information, including user identification, from presence traces.

Model understanding (Gulf of evaluation). There is a general agreement in the community that the explanations given to users should help them have an understanding on how the inference task and model work. In particular, the explanations should provide two levels of understanding to users. First, the explanations should help users understand individual decisions and predictions given by the model with respect to specific inputs. As so, the explanations justify the output of the model, that is providing the reason *why* this specific decision or prediction was given to the end-user. This justification should address the gap between the users' intentions and their perceived functionality of the system. In the case of our interpretability framework for personal information inference, the framework should provide an explanation justifying why the personal information was inferred. Second, the explanation should give the users a comprehensive understanding of the overall machine learning model when possible. Indeed, the explainability of deep learning algorithms is still an open problem. Most of the existing work is based on data visualization at the moment [4]. In this case, the explanations should give a conceptual representation of the model and provide an intuition of *how* it works. Therefore, an essential requirement consists in detailing the features at the basis of the learning process.

Model interactions (Gulf of execution). Once the user have a general understanding of how the model works, they should be able to control and interact with it. In particular, we believe that the explanations should address two interaction forms that the users might have with the system. The first type of interaction form aims at explaining the expected behavior of the system if users change the inputs or conditions of the model. Understanding how the system behaves depending on different inputs enables the users to learn how the system works. In the case of the interpretability framework for personal information inference, users should be able to determine which information can be inferred depending on the presence points they give as inputs. The second type of interaction form should aim at providing explanations and recommendations to the users such that they have an understanding of which inputs and conditions to change in order to achieve expected predictions or decisions. The users may have different expectations with respect to the predictions and decisions they want to change, ranging from abstract (*e.g.*, achieve an expected behavior) to concrete changes (*e.g.*, change the value of the model prediction). In the case of the interpretability framework, users may want to know how to prevent any personal information from being disclosed, which would require, for example, the framework to give recommendations on the places and times they should avoid visiting.

Privacy-preserving explanations. Whether the explanations are given to the service provider or a specific user of the service, the explanations must have different privacy levels. If the end-user is the service provider, the explanations do not have any privacy requirement, as the end-user can have unlimited access to the full digital traces of the users using the service. As so, the explanations can present any anonymized information about any user of the service. Alternatively, if the end-users are specific users of the service, the explanations should preserve the privacy of other individuals accessing it. In particular, the explanation cannot compromise or expose information about other user whose digital traces are used to train the inference models. This requirement poses a challenge, as both of the two above requirements may involve exposing a subset of traces from other users to the person who requested the explanation. A possible solution consists in obfuscation techniques that can be used to "hide" or aggregate the information about other users. In the case of the interpretability framework, the explanation that details why personal information was inferred should exclude information about other individual users. In particular, the explanation should not present a user's digital trace that can lead back to the single user who generated it.

2.2. Overview of the Interpretability Framework

We now present an overview of the interpretability framework for privacy of digital traces. We designed the framework such that it addresses the three main requirement presented in the previous section. The architecture of the framework is depicted in Figure 1 and consists of two main components detailed in the following: (i) the inference component and (ii) the explanation component.

Inference component. The inference component is in charge of inferring personal information about a user from their digital traces. The extent of digital traces generated by users is broad and includes IoT and pervasive system traces (*e.g.*, smartmeters and home appliances), presence traces (*e.g.*, Oyster-like smartcards, GPS traces collected by an application on the users' smartphones, check-ins at location made on location-based social networks), and online activity. The machine learning model of the inference component is specific to the inference task at hand and generally consists of a classification model that infers personal information such as activity and personal traits from the traces alone. Additionally, the component can further identify the most likely user who could have generated a set of points given as input. In particular, given a set of inputs, the inference task outputs predictions, which consist in a set of activities and personal traits associated with their respective likelihood of belonging to that activity or trait, as well as associated metadata, that is the set of per-user intermediate computation steps involved in the execution of the inference task. The inferred pattern or trait will be the one with the maximum likelihood.

Explanation component. The explanation component provides an interface for supporting user interactions. Users send queries of pre-defined types to this component, which in turn, translates them into tasks carried out by the inference component. The explanation component then retrieves the prediction outputs and metadata of the inference tasks and present them in an intelligible manner to the user. The explanation component should select relevant features given as inputs to the user. As we will detail in Section 2.3, the selected features depend on the classifier implemented by the inference component. In order to address the model understanding requirement, the explanation component should also detail the intermediate computation steps involved in the execution of the inference task. Furthermore, as discussed previously in Section 2.1, the explanation component must enable interactions with end-users of two broad types. The first type of end-users are service providers that have no privacy requirement and unlimited access to user digital traces. These end-users are interested in inferring specific activities or traits, or identifying specific users in the service from a subset of their past and current traces. For instance, in the case of user identification from presence traces, the end-users would be interested in identifying a user from a few spatio-temporal points. The second type of end-users are individual users of the service who only have access to the digital traces they generated. These end-users are interested in getting information about their current privacy state and recommendations on possible actions that could affect their privacy state. In particular, in the case of user identification from presence traces, users could be interested in knowing which of their previously recorded locations expose their privacy, or in recommendations to turn on or off location services based on where they are and the current privacy exposition level.

2.3. Implementation of the explanation component and feature selection

The explanation component relies rely on the input features and the related prediction metadata to present effective explanations to the user. However, presenting all the features and the metadata would be overwhelming for the user and could hinder the understanding of the model. As so, the explanation component must select the most relevant features that contribute the most to the particular inference task requested by the user. We believe that feature selection is the most suitable way for providing an explanation of the inference task. The feature selection method depend on the underlying machine learning model used for the

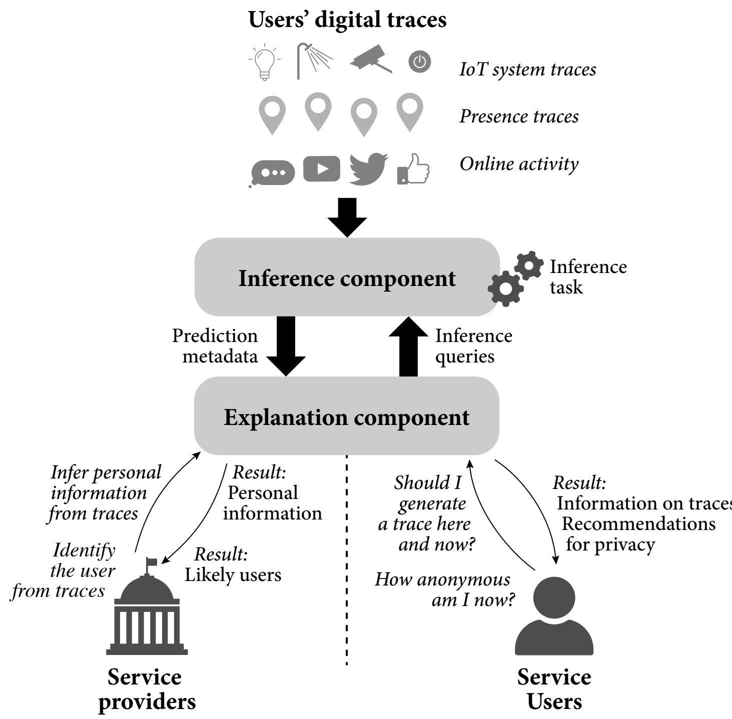


Figure 1: General conceptual architecture of the interpretability framework.

inference component and can be divided into three broad categories [7]. Each of the methods gives a score to each feature such that low-scoring features have a low relevance with respect to the other features, and should be omitted when presenting an explanation to the end-users. In Table 1 we summarize the different feature selection methods and the different classification models implemented in various the inference components of the literature we have detailed in the previous section.

3 | *Case Study*

In this section, we present a case study concerning an application of our interpretability framework to the problem of users' identification from presence data in order to provide a support for warning a user about the privacy implications of a potential check-in, *i.e.*, about their *identifiability* from a sequence of location/presence information.

We consider location information generated from smartphones and in particular check-ins from Foursquare/ Swarm, a location-based social network. As discussed, we selected this type of data given their availability and scale, but it can be considered as representative of a large class of presence data generated by IoT and pervasive systems, such wearables and public transit smartcards. Check-ins are one of the features offered by Foursquare that enable users to record the time at which they visited a location (*e.g.*, at a restaurant, in a train station or at a university). For illustrative purposes, we consider an inference component that relies on the Multinomial Naïve Bayes classifier [12]. In the following, we detail the implementation of both the inference and explanation components introduced in Section 2.2.

Table 1: Feature selection methods according to the inference component used in the literature.

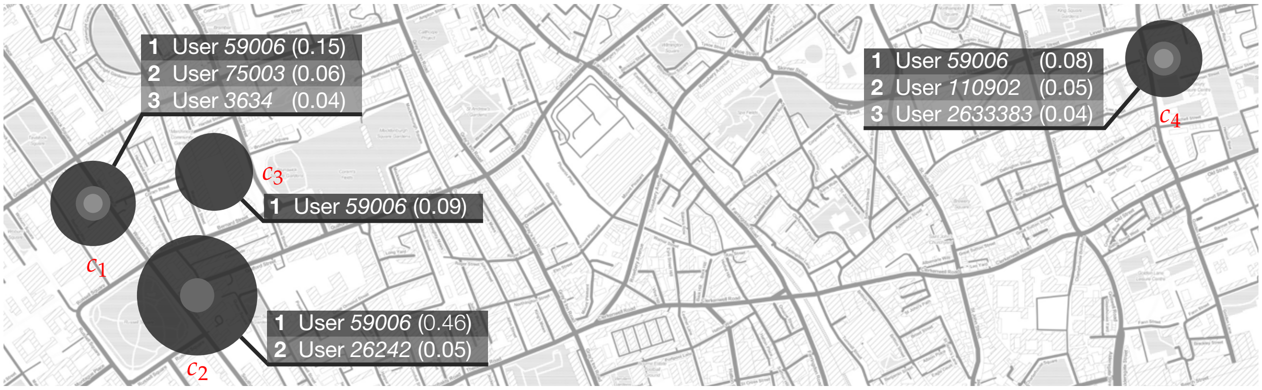
Inference component	Explanation method for inference	Potential applications
Cluster matching	Feature rank through entropy measure	Home activity inference from meter traces [8], Demographic inference from presence traces [9]
Linear regression	Feature rank through coefficients	Trait and interest inference [1]
Logistic regression	Feature ranking through the results of maximum likelihood estimation	Trait and interest inference [1]
Naïve Bayes	Feature rank through Information Gain, RelieF weights, or likelihoods	Behaviour-based identification [10], modality-based identification [11]
Multinomial Naïve Bayes	Feature rank through likelihoods and accuracy gain	Location-based identification [12]
Nearest Neighbors	Feature rank through nearest neighbor distances	Indoor location inference [13]
Linear Support Vector Machine	Feature rank through support vector weights	Behaviour-based identification [10]
Decision Trees	Feature rank through Information Gain, χ^2 test, or tree level and frequency	Behaviour-based identification [10]
Random Forest	Feature rank through Gini importance or Information Gain	Behaviour-based identification [10]

3.1. Inference Component Implementation

We first describe how the inference component identifies a user from a set of check-ins given by the end-user. The inference task will give information about the identifiability of a user. The Multinomial Naïve Bayes classifier computes the probability that a given input (*i.e.*, a set of check-ins given by the end-user) has of belonging to each output (*i.e.*, the Foursquare users). The output with the highest probability becomes the predicted label for the input (*i.e.*, the identified user). Let us consider an example training set that comprises the check-ins generated by three users A, B, and C. If the classifier computes that a given set of check-ins has 60% probability of belonging to user A, 30% probability of belonging to user B, and 10% of belonging to user C, the output of the algorithm will be that A is the most probable identity for the user.

We now provide a formal summary of the model presented in [12]. We use u to indicate a single user in the collection of all active Foursquare users U, and c_i to represent a single check-in as part of the set of check-ins $C = \{c_1, \ldots, c_n\}$ given by an end-user. Each check-in c_i is associated with a location identifier l_i and a user identifier u_i. A training set containing the check-ins associated with the Foursquare user identifiers is fed to the classifier, while the set of check-ins provided by the end-user does not contain any user identifier, as the inference task will consist in determining the correct user identifier. In other words, in our case study, the goal of the inference task is to identify the user $u^\star$ that maximizes the following product of the likelihood that each check-in belongs to the user:

$$u^\star = \arg\max_{u \in U} P(u) \prod_{i=1}^{n} P(c_i \mid u). \tag{1}$$



User *59006* was identified from the set of check-ins $\{c_1, c_2, c_3, c_4\}$.

Per-factor likelihood. The most discriminating check-in locations to identify User *59006* are:

- Check-in c_2 because $p_{c_2} = 0.46$
- Check-in c_1 because $p_{c_1} = 0.15$
- Check-in c_3 because $p_{c_3} = 0.09$
- Check-in c_4 because $p_{c_4} = 0.08$

Likelihood product. With a prior of 0.00022, User *59006* was identified from the set of check-ins $\{c_1, c_2, c_3, c_4\}$ with the top likelihood product of 1.11×10^{-7}.

Figure 2: Example of explanation of an identification task. The instance consists in an explanation of the inference task of user *59006* from the four check-ins $\{c_1, c_2, c_3, c_4\}$ represented in the figure.

3.2. Explanation Component Implementation

We consider the case of potential identification from a set of n unlabeled check-ins $C = \{c_1, \ldots, c_n\}$ using the interpretability framework. The check-in set is sent to the explanation component, which queries the inference component to determine the most likely Foursquare users who could have generated the check-ins. More specifically, the inference component transmits the individual check-in likelihoods $P(c_i \mid u)$ for each check-in c_i of C and each user u in the training set. Using this data, the explanation component then presents the explanation of the result of the inference task to the end-user. However, the component must limit the amount of information to present, as too much information can be overwhelming and irrelevant for the explanation. For instance, mentioning the users who did not check-in at a location would not be relevant to the end-user. Similarly, when a significant number of Foursquare users have checked-in at a location, it can be overwhelming for the end-user to present information about *all* users. As so, in our explanation component prototype, we limit the information presented to the end-user to the top three Foursquare users that have the best check-in likelihood at each location given as inputs. We represent an example explanation presented by the framework to the end-user in Figure 2. As illustrative dataset, we consider the Foursquare check-ins collected at a world-wide scale during a period of about 18 months from April 2012 to September 2013 by Yang *et al.* [14]. We further restrict the user check-in locations to all the boroughs of the greater London area. This results in 184,557 check-ins of 9,687 users within London at 28,218 different locations.

We choose to present a visual representation of the check-ins, as presence data can be easily plotted on a map. In this example, the end-user wishes to perform an identification task of a Foursquare user from the set of check-ins $C = \{c_1, c_2, c_3, c_4\}$ represented in the figure. With the set of check-ins, the inference component (correctly) identifies user *59006* among the user traces contained in the training set. The figure depicts the explanation provided by the corresponding component to the end-user. The explanation in this case consists in the top three check-in likelihoods of the users for each location. For instance, at check-in location c_1, user *59006* has the best check-in likelihood of $p_{c_1,u_{59006}} = 0.15$, followed by users *75003* and *3634* with check-in likelihood of $p_{c_1,u_{75003}} = 0.06$ and $p_{c_1,u_{3634}} = 0.04$, respectively. In fact, user *59006* checked-in seven times at location c_1 out of 35 check-ins in total, which corresponds to a check-in likelihood of $p_{c_1} = 0.15$ (including the smoothing parameter α). We also omitted the information about users that did not check in at the places, as this information is not relevant to the explanation process.

An explanation of the specific inference task is provided to the user by presenting the relevant features. Information about the machine learning model itself is also provided by detailing the intermediate computation steps involved in the classification. The explanation component should allow users to interact with the inference component in order to get additional information regarding the identifiability of the traces they generated and shared with the service.

The explanation should provide a clear recommendation as to whether the user should check-in or not at that location; in particular, it must then rely on an estimate of the relevance of the new check-in with respect to the current identifiability of the user. As we detail in Section 2.3, the relevance of a feature, here a check-in, can be estimated through feature selection methods. In our case the inference component is based on a Multinomial Naïve Bayes classifier and the best suited feature selection method consists of a wrapper that computes the accuracy gain of the inference task with and without the new check-in.

4 | *Open Research Challenges and Outlook*

In this section, we detail the different challenges related to the interpretation and explanation of machine learning technique for inference of personal information that expose the privacy of users from the digital traces they generate through the use and the interaction with IoT and pervasive systems.

Information selection. By default, the model is seen as a black box with a set of inputs and outputs. Predictions alone and metrics based on them do not suffice to characterize and explain the model [4]. While this information is limited, knowing the internals of the model would allow system designers to present relevant information that was learned by the model during its training. The challenge is to determine the amount of information an explanation component should present to users so that they have an understanding of how the model works. As noted by Kulesza *et al.* in [15], how much information to present to the user remains an open question. This includes determining which features to select and present to the users in order to avoid to give an overwhelming quantity of information, which could result in very complex set of explanations that might not be intelligible. Indeed, while extreme simplicity is not acceptable for users, giving too much information to the user can be overwhelming and will decrease the *quality* of the explanation.

Level of expertise of the end-users. The explanation presented to the users must take into account the level of expertise of the user. In particular, the quantity and the complexity of information contained in the explanation should depend on the expertise of the user. As so, an expert user will require extensive explanations with more information, whereas a non-expert user will need simple explanations with a limited amount of information. The challenge consists in determining the appropriate amount of information and the level of detail to present to the user. A possible experimental strategy is to conduct user studies in order to allow researchers and practitioners to determine the necessary and sufficient amount of information. The definition of the methodology of these studies is another open challenge *per se* the interpretable machine learning community as discussed below. The challenge is that user studies might be very specific to the particular inference techniques and it might be difficult to extract general conclusions from them [15].

Definition of privacy-preserving interpretation and anonymity. The explanation given to a user cannot compromise or expose information about other individuals whose data is contained in the training set of the machine learning model. When presenting an explanation, sensitive information about other users may be disclosed by the information given in the explanation. The explanations can rely on obfuscation techniques to hide information about other users when presenting an explanation. Obfuscation techniques include aggregation methods to guarantee a k-anonymity with respect to this information of other users [16]. With a k-anonymity guarantee, users cannot be distinguished from at least $k - 1$ individual users whose information appear in the explanation. Additionally, presenting recommendations to the users so that they can take future actions to better protect their privacy requires presenting explanations that contain information about other individuals. This task is particularly challenging, as it requires presenting the general patterns associated to an individual, without revealing the specific patterns associated to other users.

Data shift and recommendations. Giving explanations to the users allow them to understand why their privacy was exposed. With recommendations, users can take actions on their future behavior to further protect their privacy. This task is challenging for the following reasons. First, the user must trust the system to follow the recommendation. As we have seen earlier in this section, trust can be gained with effective explanations of the inference task. Second, the recommendation given to a specific user depends on the expected future behavior of the other users, as well as whether they follow the recommendations that they received. As a result, giving recommendations to users may change their behavior and routines, which may in turn invalidate the current training set. This problem is often referred to as the data shift problem [17]. A solution would consist in training the model again using recently generated traces, but this might not be always possible in practical situations for a variety of reasons.

Evaluation of the explanations. When giving automatically generated explanations to users, we need a methodology to evaluate the interpretability of the explanation in terms of effectiveness and fidelity of the inference model with respect to the requirements listed in Section 2.1 (*i.e.*, the quality of the model understanding and interaction, as well as the privacy respect). Since interpretability depends on a human judgement, it is subjective and depends on the background and level of expertise of the evaluator. Qualitative evaluation is then necessary and can be performed through user studies or surveys. In this case the challenge is to design a methodology to test and assess the subjective comprehensibility of the explanation with user studies by choosing the right set of users, including expert users and non-expert users. In particular, it is important to determining the performance metrics used to reliably evaluate the explanation. For instance, trust in the explanation is a common metric that is highly subjective. While qualitative evaluations are necessary, an interesting research question is how to perform systematic quantitative evaluation of explanations given target performance metrics, without the involvement of users [2].

Presenting the explanation to the end-users. The explanation component is in charge of delivering effective explanations to the end-user. However, the explanations can have various types of representation, which are usually classified into three broad categories, namely visual explanations, textual explanations, or explanations by examples [4]. Visual or textual representations depend on the amount of data to present and the expertise of the end-user. Intuitively, an expert end-user would prefer more complete and textual explanations, while a non-expert end-user would rather have visual and interactive explanations. The challenge is then to determine the best suited representation of the explanations to deliver to the end-users.

We believe that addressing these research challenges will be possible only through the collaboration of researchers from different communities, including, but not limited to, pervasive computing, human-computer interaction, data visualization and machine learning.

References

[1] M. Kosinski, D. Stillwell, and T. Graepel, "Private traits and attributes are predictable from digital records of human behavior," *Proceedings of the National Academy of Sciences*, vol. 110, no. 15, pp. 5802–

5805, 2013.

[2] F. Doshi-Velez and B. Kim, "A roadmap for a rigorous science of interpretability," *arXiv preprint arXiv:1702.08608*, 2017.

[3] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *ACM CHI*, (Boston, MA, USA), Apr. 2009.

[4] Z. C. Lipton, "The mythos of model interpretability," in *ICML Workshop on Human Interpretability in Machine Learning*, (Sydney, Australia), 2017.

[5] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *ACM UbiComp*, (Orlando, FL, USA), pp. 195–204, Sept. 2009.

[6] D. A. Norman and S. W. Draper, *User centered system design*. L. Erlbaum Associates Inc., 1986.

[7] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[8] M. A. Lisovich, D. K. Mulligan, and S. B. Wicker, "Inferring personal information from demand-response systems," *IEEE Security & Privacy*, vol. 8, no. 1, 2010.

[9] K. Mohamed, E. Côme, J. Baro, and L. Oukhellou, "Understanding passenger patterns in public transit through smart card and socioeconomic data," in *UrbComp*, (Seattle, WA, USA), Sept. 2014.

[10] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *ACM SIGKDD*, (Chicago, IL, USA), Aug. 2013.

[11] W. Shi, J. Yang, Y. Jiang, F. Yang, and Y. Xiong, "Senguard: Passive user identification on smartphones using multiple sensors," in *IEEE WiMob*, (Shanghai, China), Oct. 2011.

[12] L. Rossi and M. Musolesi, "It's the way you check-in: identifying users in location-based social networks," in *ACM COSN*, (Dublin, Ireland), Oct. 2014.

[13] L. M. Ni, Y. Liu, Y. C. Lau, and A. P. Patil, "LANDMARC: indoor location sensing using active RFID," *Wireless networks*, vol. 10, no. 6, pp. 701–710, 2004.

[14] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, p. 30, 2016.

[15] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *IEEE VL/HCC*, (San Jose, CA, USA), Sept. 2013.

[16] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," tech. rep., SRI International, 1998.

[17] J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, *Dataset shift in machine learning*. The MIT Press, 2009.

Authors' Biographies

Benjamin Baron is a Postdoctoral Research Fellow at University College London. His interests include digital traces analytics, mobile phone data modeling and security&privacy. Dr Baron has a Ph.D. in Computer Science from UPMC Sorbonne Universités.

Mirco Musolesi is a Reader in Data Science at University College London and Turing Fellow at the Alan Turing Institute. His research interests include ubiquitous computing, intelligent systems and security&privacy. Dr Musolesi has a Ph.D. in Computer Science from University College London.