# An Information-Theoretic Approach to Personalized Explainable Machine Learning

Alexander Jung and Pedro H. J. Nardelli

***Abstract*—Automated decision making is used routinely throughout our every-day life. Recommender systems decide which jobs, movies, or other user profiles might be interesting to us. Spell checkers help us to make good use of language. Fraud detection systems decide if a credit card transactions should be verified more closely. Many of these decision making systems use machine learning methods that fit complex models to massive datasets. The successful deployment of machine learning (ML) methods to many (critical) application domains crucially depends on its explainability. Indeed, humans have a strong desire to get explanations that resolve the uncertainty about experienced phenomena like the predictions and decisions obtained from ML methods. Explainable ML is challenging since explanations must be tailored (personalized) to individual users with varying backgrounds. Some users might have received university-level education in ML, while other users might have no formal training in linear algebra. Linear regression with few features might be perfectly interpretable for the first group but might be considered a black-box by the latter. We propose a simple probabilistic model for the predictions and user knowledge. This model allows to study explainable ML using information theory. Explaining is here considered as the task of reducing the "surprise" incurred by a prediction. We quantify the effect of an explanation by the conditional mutual information between the explanation and prediction, given the user background.**

## I. Introduction

Machine learning (ML) methods compute predictions for quantities of interest based on a statistical analysis of large amounts of historical data [2], [10], [14]. These methods are routinely used to power many services within our everyday-life. ML methods power recommendation systems that decide what job ads or which other user profiles could be interesting to us [17], [26]. Recent breakthroughs in ML, such as in image or text processing [8], also holds the promise of boosting the level of automation in domains which currently rely mainly on human labour or manual design [7].

A key challenge for the successful and ethically sound deployment of ML methods to critical application domains is the (lack of) explainability of its predictions [9], [12], [19], [25]. Explanations of predictions, which are used for decisions that affect humans, are increasingly becoming a legal obligation [25]. Beside legal aspects, it also seems that humans have a basic need for understanding decision making processes [15], [16].

One reason why explainable ML is challenging is that (good) explanations must be tailored to the knowledge of individual users ("explainee"). In general, for a particular prediction, there is no unique explanation that serves equally well a large group of heterogeneous users. Thus, achieving explainable ML would be easier for applications involving a homogenous group of users, like graduate students in a university program.

Large-scale applications as, for instance, recommendation systems for video streaming providers typically involve users with very different backgrounds, which can range from graduate studies in ML-related fields to users with no formal training in linear algebra. While linear models involving few hand-crafted features might be viewed as interpretable for the former group it might be considered a "black-box" for the latter group of users.

This contribution studies explainable ML within information theory by using a probabilistic model for the data and user background. Loosely speaking, we model the effect of providing an explanation for a prediction as a reduction of the "surprise" incurred by a prediction to the user. This qualitative interpretation of explaining a prediction leads naturally to measuring the quantitative effect of explanations via (conditional) mutual information (MI) between the explanation and the prediction, given the user background (see Section II).

Our approach is different from existing work on explainable ML in the sense that we explicitly model the specific knowledge of each individual user. In contrast, most existing methods for explainable ML do not make any assumption about the end-user and her background knowledge.

Explainable ML methods can be roughly divided into two groups. The first group of methods uses models that are considered as intrinsically interpretable, like linear regression or small decision trees. The second group of methods, referred to as model-agnostic methods, probe an ML method by perturbing the features of the data point.

The most straightforward approach to explainable ML methods is to use models that are considered to be intrinsically interpretable. Such methods include linear models, decision trees and artificial neural networks [1], [9], [21]. Explaining the predictions obtained from such intrinsically interpretable models merely amounts to specifying the model parameters, such as the weights $w_i$ of a linear predictor $h(\mathbf{x}) = \sum_i w_i x_i$, or the feature-wise thresholds used in decision trees [10].

Interpretable models offer an intuitive decomposition of its predictions into a combination of elementary properties of a data point. Defining elementary properties of a data point via the activations of a (deep) neural network renders those models also interpretable (see [21]).

Explainable models for sequential decision making have been studied in [18], where the authors obtain an explainable multi-

AJ is with the Department of Computer Science, Aalto University, Finland. PHJN is with Lappeenranta-Lahti University of Technology, Finland. This work is partly supported by Academy of Finland via: (a) ee-IoT project n.319009, (b) FIREMAN consortium CHIST-ERA/n.326270, and (c) EnergyNet Research Fellowship n.321265/n.328869.

armed bandit model by using the choice for the action space as the explanation. An explanation can be obtained by notifying the user that only previously purchased items are recommended. In contrast to [18], our approach uses a probabilistic model for the user background to compute personalized explanations that are optimal in a precise (information-theoretic) sense.

A second group of explainable ML, referred to as model agnostic methods, is based on constructing explanations by probing a predictor as a black box [9], [24]. These methods aim at locally approximating black box models by simpler and interpretable models, such as linear models or shallow decision trees [24].

Our approach is also model agnostic as it only requires the statistical distributions of the model prediction. However, in contrast to most model agnostic explainable ML, we do not use local approximations to explain a black box method. Instead, we use a probabilistic model for the predictions and user knowledge.

We frame explainable ML within a probabilistic model for ML predictions and user knowledge. This allows to capture the act of explaining a prediction using information-theoretic concepts. The act of explaining provides the user additional information about the prediction delivered by some (arbitrary) ML method.

Information theory has already been used for learning optimal explanations [3]. In a similar spirit, we also use MI to guide the learning of instance-wise explanations. However, in contrast to [3], we also model the effect of the user background on the information provided by an explanation. In a nutshell, while [3] uses unconditional MI between explanations and predictions, we use the conditional MI given the user knowledge (see Section III).

**Outline and Contribution.** In Section II, we propose a simple probabilistic model for the features, prediction and user summary of a data point. This probabilistic model allows to quantify the effect of explanations via the conditional between the explanation and the model prediction, given the user background.

Our main contribution is the formulation of an information-theoretic concept of optimal personalized explanations. As discussed in Section III), we construct (information-theoretically) optimal personalized explanations by maximizing the conditional MI between explanation and predictions, when conditioning on the user summary of a data point. To the best of our knowledge, we present the first information-theoretic approach to personalized explainable ML.

A simple algorithm for computing optimal explanation given the model predictions and user summaries based on i.i.d. samples is presented in Section IV. The proposed algorithm allows to construct personalized explanations that are optimal in an information-theoretic sense.

## II. Problem Setup

We consider a supervised ML problem involving data points with features $\mathbf{x} = \left(x_1, \ldots, x_n\right)^T \in \mathbb{R}^n$ and label $y \in \mathbb{R}$. Given some labelled training data

$$\left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right), \ldots, \left(\mathbf{x}^{(m)}, y^{(m)}\right), \tag{1}$$

ML methods typically learn a predictor (map)

$$h(\cdot) : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto \hat{y} = h(\mathbf{x}) \tag{2}$$

by requiring $\hat{y}^{(i)} \approx y^{(i)}$ [2], [10], [14].

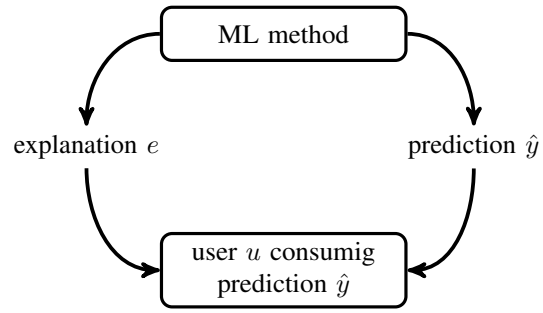


Fig. 1. An explanation $e$ provides additional information $I(\hat{y}, e|u)$ to a user $u$ about the prediction $\hat{y}$.

After learning a predictor $\hat{y} = h(\mathbf{x})$, it is applied to new data points yielding the prediction $\hat{y} = h(\mathbf{x})$. In may application, the prediction $\hat{y}$ is then delivered to a human user. The user can be the subscriber of a streaming service [6], a dermatologist [5] or a city planner [27].

Each user has typically some conception or model for the relation between features $\mathbf{x}$ and label $y$ of a data point. Based on the user background, she has some understanding of a data point with features $\mathbf{x}$.

Our approach to explainable ML is based on modelling the user understanding of a data point by some summary $u \in \mathbb{R}$. The summary is obtained by a (stochastic) map from the features $\mathbf{x}$ of a data point. We will focus on summaries being obtained by a deterministic map

$$u(\cdot) : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto u := u(\mathbf{x}). \tag{3}$$

However, our approach also covers stochastic maps characterized by a conditional probability distribution $p(u|\mathbf{x})$.

The (user-specific) quantity $u$ represents the understanding of the specific properties of the data point given the user knowledge (modelling assumptions). We interpret $u$ as a "summary" of the data point based on its features $\mathbf{x}$ and the intrinsic modelling assumptions of the user.

Let us illustrate the concept of the user summary $u$ as a means to represent user knowledge (or background) by two particular choices for $u$. First, the user summary could be the prediction obtained from a simplified model, such as linear regression using few features that the user anticipates as being relevant. Another example for a user summary $u$ could be a higher-level feature, such as eye spacing in facial pictures [13].

We formalize the act of explaining a prediction $\hat{y} = h(\mathbf{x})$ as presenting some additional quantity $e$ to the user. This "explanation" $e$ can be any quantity that helps the user to understand the prediction $\hat{y}$, given her understanding $u$ of the data point. Loosely speaking, the explanation $e$ contributes to resolving the uncertainty of the user $u$ about the prediction $\hat{y}$ [15].

For the sake of exposition, our focus will be on explanations

obtained via a deterministic map

$$e(\cdot) : \mathbb{R}^n \to \mathbb{R} : \mathbf{x} \mapsto e := e(\mathbf{x}), \tag{4}$$

from the features $\mathbf{x}$ of a data point. However, our approach can be generalized without difficulty to handle explanations obtained by a (stochastic) map. In the end, we only require the specification of the conditional probability distribution $p(e|\mathbf{x})$.

Explanations can be constructed in quite different ways. An explanation could be a subset of features of a data point (see [23] and Section III). More generally, explanations could be obtained from simple local statistics (averages) of features that are considered related, such as near-by pixels in an image or consecutive samples of an audio signal. Instead of individual features, carefully chosen data points can also serve as an explanation [18], [24].

To obtain comprehensible explanations that can be computed efficiently, we must typically restrict the space of possible explanations to a small subset $\mathcal{F}$ of maps (10). This is conceptually similar to the restriction of the space of possible predictor functions in a ML method to a small subset of maps which is known as the hypothesis space.

We consider data points as independent and identically distributed (i.i.d.) realizations of a random variable with fixed underlying probability distribution $p(\mathbf{x}, y)$. Modelling the data point as random implies that the user summary $u$, prediction $\hat{y}$ and explanation $e$ are also random variables. The joint distribution $p(u, \hat{y}, e, \mathbf{x}, y)$ conforms with the Bayesian network [22] (depicted in Figure 2) since

$$p(u, \hat{y}, e, \mathbf{x}, y) = p(u|\mathbf{x}) \cdot p(e|\mathbf{x}) \cdot p(\hat{y}|\mathbf{x}) \cdot p(\mathbf{x}, y). \tag{5}$$

We measure the amount of additional information provided by an explanation $e$ for a prediction $\hat{y}$ to some user $u$ via the conditional MI [4, Ch. 2 and 8]

$$I(e; \hat{y}|u) := \mathrm{E}\left\{ \log \frac{p(\hat{y}, e|u)}{p(\hat{y}|u)p(e|u)} \right\}. \tag{6}$$

The conditional MI $I(e; \hat{y}|u)$ can also be interpreted as a measure for the amount by which the explanation $e$ reduces the uncertainty about the prediction $\hat{y}$ which is delivered to some user $u$. Thus, constructing explanations via solving (6) conforms with the apparent human need to understand observed phenomena, such as the predictions from a ML method [15].

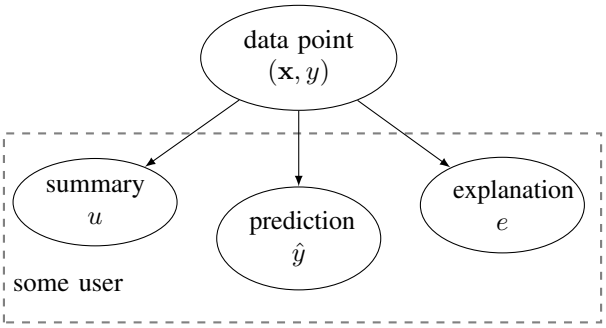


Fig. 2. A simple probabilistic model for explainable ML.

## III. Optimal Explanations

Capturing the effect of an explanation using the probabilistic model (6) offers a principled approach to computing an optimal explanation $e$. We require the optimal explanation $e^*$ to maximize the conditional MI (6) between the explanation $e$ and the prediction $\hat{y}$ conditioned on the user summary $u$ of the data point.

Formally, an optimal explanation $e^*$ solves

$$I(e^*; \hat{y}|u) = \sup_{e \in \mathcal{F}} I(e; \hat{y}|u). \tag{7}$$

The choice for the subset $\mathcal{F}$ of valid explanations offers a trade-off between comprehensibility, informativeness and computational cost incurred by an explanation $e^*$ (solving (7)).

The maximization problem (7) for obtaining optimal explanations is similar to the approach in [3]. However, while [3] uses the unconditional MI between explanation and prediction, (7) involves the conditional MI given the user summary $u$.

Let us illustrate the concept of optimal explanations (7) using a linear regression method. We model the features $\mathbf{x}$ as a realization of a multivariate normal random vector with zero mean and covariance matrix $\mathbf{C}_x$,

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x). \tag{8}$$

The predictor and the user summary are linear functions of the features,

$$\hat{y} := \mathbf{w}^T \mathbf{x}, \text{ and } u := \mathbf{v}^T \mathbf{x}. \tag{9}$$

We construct explanations via subsets of individual features $x_i$ that are considered most relevant for a user to understand the prediction $\hat{y}$ (see [21, Definition 2] and [20]). Thus, we consider explanations of the form

$$e := \{x_i\}_{i \in \mathcal{E}} \text{ with some subset } \mathcal{E} \subseteq \{1, \ldots, n\}. \tag{10}$$

The complexity of an explanation $e$ is measured by the number $|\mathcal{E}|$ of features that contribute to it. We limit the number of features contributing to an explanation by a fixed (small) sparsity level,

$$|\mathcal{E}| \leq s (\ll n). \tag{11}$$

Modelling the feature vector $\mathbf{x}$ as Gaussian (8) implies that the prediction $\hat{y}$ and user summary $u$ obtained from (9) is jointly Gaussian for a given $\mathcal{E}$ (10). Basic properties of multivariate normal distributions [4, Ch. 8], allow to develop (7) as

$$\begin{aligned} \max_{\substack{\mathcal{E} \subseteq \{1,\ldots,n\} \\ |\mathcal{E}| \leq s}} & I(e; \hat{y}|u) \\ &= h(\hat{y}|u) - h(\hat{y}|u, \mathcal{E}) \\ &= (1/2) \log \mathbf{C}_{\hat{y}|u} - (1/2) \log \det \mathbf{C}_{\hat{y}|u,\mathcal{E}} \\ &= (1/2) \log \sigma^2_{\hat{y}|u} - (1/2) \log \sigma^2_{\hat{y}|u,\mathcal{E}}. \end{aligned} \tag{12}$$

Here, $\sigma^2_{\hat{y}|u}$ denotes the conditional variance of the prediction $\hat{y}$, conditioned on the user summary $u$. Similarly, $\sigma^2_{\hat{y}|u,\mathcal{E}}$ denotes the conditional variance of $\hat{y}$, conditioned on the user summary $u$ and the subset $\{x_r\}_{r \in \mathcal{E}}$ of features. The last step in (12) follows from the fact that $\hat{y}$ is a scalar random variable.

The first component of the last expression in (12) does not depend on the choice $\mathcal{E}$ for the explanation $e$ (see (10)). Therefore, the optimal choice $\mathcal{E}$ solves

$$\sup_{|\mathcal{E}|\leq s} -(1/2)\log \sigma^2_{\hat{y}|u,\mathcal{E}}. \tag{13}$$

The maximization (13) is equivalent to

$$\inf_{|\mathcal{E}|\leq s} \sigma^2_{\hat{y}|u,\mathcal{E}}. \tag{14}$$

In order to solve (14), we relate the conditional variance $\sigma^2_{\hat{y}|u,\mathcal{E}}$ to a particular decomposition

$$\hat{y} = \alpha u + \sum_{i\in\mathcal{E}} \beta_i x_i + \varepsilon. \tag{15}$$

For an optimal choice of the coefficients $\alpha$ and $\beta_i$, the variance of the error term in (15) is given by $\sigma^2_{\hat{y}|u,\mathcal{E}}$. Indeed,

$$\min_{\alpha,\beta_i\in\mathbb{R}} \mathrm{E}\big\{\big(\hat{y} - \alpha u - \sum_{i\in\mathcal{E}} \beta_i x_i\big)^2\big\} = \sigma^2_{\hat{y}|u,e}. \tag{16}$$

Inserting (16) into (14), an optimal choice $\mathcal{E}$ (of feature) for the explanation of prediction $\hat{y}$ to user $u$ is obtained from

$$\inf_{|\mathcal{E}|\leq s} \min_{\alpha,\beta_i\in\mathbb{R}} \mathrm{E}\big\{\big(\hat{y} - \alpha u - \sum_{i\in\mathcal{E}} \beta_i x_i\big)^2\big\} \tag{17}$$

$$= \min_{\|\boldsymbol{\beta}\|_0\leq s} \mathrm{E}\big\{\big(\hat{y} - \alpha u - \boldsymbol{\beta}^T\mathbf{x}\big)^2\big\}. \tag{18}$$

An optimal subset $\mathcal{E}_{\rm opt}$ of features defining the explanation $e$ (10) is obtained from any solution $\boldsymbol{\beta}_{\rm opt}$ of (18) via

$$\mathcal{E}_{\rm opt} = \mathrm{supp}\,\boldsymbol{\beta}_{\rm opt}. \tag{19}$$

## IV. A Simple XML Algorithm

Under a Gaussian model (8) for the features of data points, Section III shows how to construct optimal explanations via the (support of the) solutions $\boldsymbol{\beta}_{\rm opt}$ of the sparse linear regression problem (18).

In order to obtain a practical algorithm for computing (approximately) optimal explanations (19), we need to approximate the expectation in (18) with an empirical average over i.i.d. samples $\big(\mathbf{x}^{(i)}, \hat{y}^{(i)}, u^{(i)}\big)$ of features, predictions and user summaries. This results in Algorithm 1.

**Algorithm 1** XML Algorithm

**Input:** explanation sparsity $s$, training samples $\big(\mathbf{x}^{(i)}, \hat{y}^{(i)}, u^{(i)}\big)$ for $i = 1,\ldots,m$

1: compute $\widehat{\boldsymbol{\beta}}$ by solving

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\|\boldsymbol{\beta}\|_0\leq s} \sum_{i=1}^m \big(\hat{y}^{(i)} - \alpha u^{(i)} - \boldsymbol{\beta}^T\mathbf{x}^{(i)}\big)^2 \tag{20}$$

**Output:** feature set $\widehat{\mathcal{E}} := \mathrm{supp}\widehat{\boldsymbol{\beta}}$

Note that Algorithm 1 is interactive since the user has to provide samples $u^{(i)}$ of its summary for the data points with features $\mathbf{x}^{(i)}$. Based on the user input $u^{(i)}$, for $i = 1,\ldots,m$, Algorithm 1 learns an optimal subset $\mathcal{E}$ of features (10) that are used for the explanation of predictions.

The sparse regression problem (20) becomes intractable for large feature length $n$. However, if the features are weakly correlated with each other and the user summary $u$, the solutions of (20) can be found by convex optimization. Indeed, for a wide range of settings, sparse regression (20) can be solved via a convex relaxation, known as the least absolute shrinkage and selection operator (Lasso) [11],

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^n} \sum_{i=1}^m \big(\hat{y}^{(i)} - \alpha u^{(i)} - \boldsymbol{\beta}^T\mathbf{x}^{(i)}\big)^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{21}$$

We have already a good understanding of choosing the Lasso parameter $\lambda$ in (21) such that its solutions coincide with the solutions of (20) (see, e.g., [11]).

## V. Numerical Experiments

We verify the ability of Algorithm 1 to provide explainable ML using a computer vision application. In particular, we consider data points representing square patches of a greyscale aerial photograph of Helsinki city area.[1] The goal is to predict the greyscale value $y$ of the center ("target") pixel. In order to predict the greyscale value of the $i$th pixel $y^{(i)}$ we use the greyscale values $x_j^{(i)}$ of close-by pixels $j \in \mathcal{P}^{(i)}$. As depicted in Figure 3, the neighbourhood $j \in \mathcal{P}^{(i)}$ is constituted by two rectangular areas that are adjacent to pixel $i$. A user having

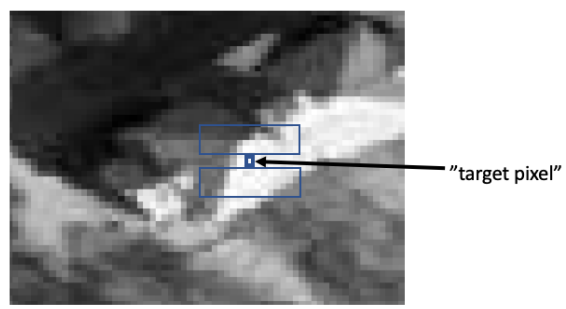


Fig. 3. The greyscale level of a particular target pixel within an aerial photograph can be predicted based on the greyscale values of nearby pixels within adjacent rectangles (indicated).

some prior experience in processing natural images might consider the average greyscale value

$$u^{(i)} = (1/|\mathcal{P}^{(i)}|) \sum_{j\in\mathcal{P}^{(i)}} x_j^{(i)} \tag{22}$$

as a reasonable summary of the features $\mathbf{x}^{(i)} := \{x_j^{(i)}\}_{j\in\mathcal{P}^{(i)}}$. We refer to the Python notebook https://github.com/alexjungaalto/ResearchPublic/blob/master/itxml.ipynb for the results of the experiments.

## VI. Conclusion

We have introduced a simple probabilistic model for the predictions of a ML method and the user background. The user background is represented by a summary of the features of a data point. The effect of an explanation is measured by the conditional MI between prediction and explanation, given the user summary of a data point.

[1]The data is freely available via the online map service https://kartta.hel.fi/

## References


[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] J. Chen, L. Song, M.J. Wainwright, and M.I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *Proc. 35th Int. Conf. on Mach. Learning*, Stockholm, Sweden, 2018.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New Jersey, 2 edition, 2006.

[5] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 2017.

[6] C.A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *Association for Computing Machinery*, 6(4), January 2016.

[7] N.J. Goodall. Can you program ethics into a self-driving car? *IEEE Spectrum*, 53(6):28–58, June 2016.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[9] H. Hagras. Toward human-understandable, explainable ai. *Computer*, 51(9):28–36, Sep. 2018.

[10] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, USA, 2001.

[11] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity. The Lasso and its Generalizations*. CRC Press, 2015.

[12] A. Holzinger. Explainable AI (ex-AI). *Informatik Spektrum*, 41:138–143, April 2018.

[13] K. Jeong, J. Choi, and G. Jang. Semi-local structure patterns for robust face detection. *IEEE Sig. Proc. Letters*, 22(9), 2015.

[14] A. Jung. Components of machine learning: Binding bits and flops. *arXiv preprint https://arxiv.org/pdf/1910.12387.pdf*, 2019.

[15] J. Kagan. Motives and development. *Journal of Personality and Social Psychology*, 22(1):51–66, 1972.

[16] A.W. Kruglanski and D.M. Webster. Motivated closing of the mind. *Psychol. Rev.*, 103(2), 1996.

[17] A. B. B. Martinez, J. J. P. Arias, A. F. Vilas, J. Garcia Duque, and M. Lopez Nores. What's on tv tonight? an efficient and effective personalized recommender system of tv programs. *IEEE Transactions on Consumer Electronics*, 55(1):286–294, 2009.

[18] J. McInerney, B. Lacker, S. Hansen, K. Higley, H. Bouchard, A. Gruson, and R. Mehrotra. Explore, exploit, and explain: personalizing explainable recommendations with bandits. In *Proceedings of the 12th ACM Conference on Recommender Systems*, 2018.

[19] B.D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, and L. Floridi. The ethics of algorithms: Mapping the debate. *Big Data and Society*, 3(2), 2016.

[20] C. Molnar. *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*. [online] Available: https://christophm.github.io/interpretable-ml-book/., 2019.

[21] G. Montavon, W. Samek, and K. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[22] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 1988.

[23] M.T. Ribeiro, S. Singh, and C. Guestrin. "Why should i trust you?": Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD*, pages 1135–1144, Aug. 2016.

[24] M.T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[25] S. Wachter, B. Mittelstadt, and L. Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, 2017.

[26] R. Wang, C. Chow, Y. Lyu, V.C.S. Lee, S. Kwong, Y. Li, and J. Zeng. Taxirec: Recommending road clusters to taxi drivers using ranking-based extreme learning machines. *IEEE Transactions on Knowledge and Data Engineering*, 30(3):585–598, 2018.

[27] X. Yang and Q. Wang. Crowd hybrid model for pedestrian dynamic prediction in a corridor. *IEEE Access*, 7, 2019.