VIME: Visual Interactive Model Explorer for Identifying Capabilities and Limitations of Machine Learning Models for Sequential Decision-Making

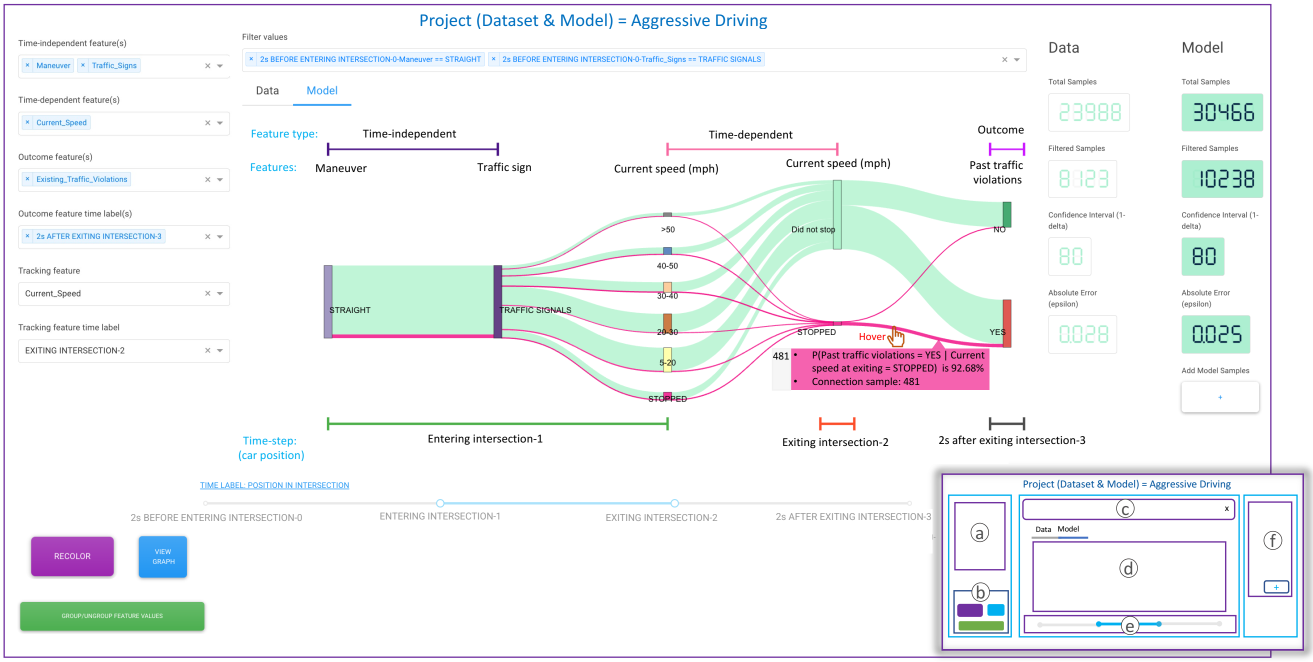
Anindya Das Antar
adantar@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Somayeh Molaei
smolaei@umich.edu
University of Michigan
Ann Arbor, Michigan, USA

Yan-Ying Chen
yan-ying.chen@tri.global
Toyota Research Institute
Los Altos, California, USA

Matthew Lee
matt.lee@tri.global
Toyota Research Institute
Los Altos, California, USA

Nikola Banovic
nbanovic@umich.edu
University of Michigan
Ann Arbor, Michigan, USA



Project (Dataset & Model) = Aggressive Driving

Time-independent feature(s): Maneuver, Traffic_Signs

Time-dependent feature(s): Current_Speed

Outcome feature(s): Existing_Traffic_Violations

Outcome feature time label(s): 2s AFTER EXITING INTERSECTION-3

Tracking feature: Current_Speed

Tracking feature time label: EXITING INTERSECTION-2

Filter values: 2s BEFORE ENTERING INTERSECTION-0-Maneuver == STRAIGHT; 2s BEFORE ENTERING INTERSECTION-0-Traffic_Signs == TRAFFIC SIGNALS

- P(Past traffic violations = YES | Current speed at exiting = STOPPED) is 92.68%
- Connection sample: 481

	Data	Model
Total Samples	23988	30466
Filtered Samples	8123	10238
Confidence Interval (1-delta)	80	80
Absolute Error (epsilon)	0.028	0.025

Figure 1: VIME interface showing a "what-if" scenario from a sequential ML model [15] that detects aggressive driving behaviors: a) a feature selection panel to select and visualize a subset of relevant features, b) a control panel to modify model output preferences, c) filtered feature values, d) interactive Sankey visualization showing a sequence of selected time-independent and time-dependent inputs and outputs, e) a range slider to zoom in and out of specific sequence timesteps, and f) sample size determination panel. The "what-if" scenario shows that the model correctly predicts that drivers who stop in the middle of an intersection are likely aggressive drivers with past traffic violations, with a 92.68% likelihood. It also shows that the model wrongly predicts that drivers can decelerate from over 50 mph to a full stop within the short length of an intersection.

ABSTRACT

Ensuring that Machine Learning (ML) models make correct and meaningful inferences is necessary for the broader adoption of such models into high-stakes decision-making scenarios. Thus, ML model engineers increasingly use eXplainable AI (XAI) tools to investigate the capabilities and limitations of their ML models before deployment. However, explaining sequential ML models, which make a series of decisions at each timestep, remains challenging. We

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

UIST '24, October 13–16, 2024, Pittsburgh, PA, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0628-8/24/10
https://doi.org/10.1145/3654777.3676323

present Visual Interactive Model Explorer (*VIME*), an XAI toolbox that enables ML model engineers to explain decisions of sequential models in different "what-if" scenarios. Our evaluation with 14 ML experts, who investigated two existing sequential ML models using VIME and a baseline XAI toolbox to explore "what-if" scenarios, showed that VIME made it easier to identify and explain instances when the models made wrong decisions compared to the baseline. Our work informs the design of future interactive XAI mechanisms for evaluating sequential ML-based decision support systems.

CCS CONCEPTS

• **Human-centered computing → Interactive systems and tools**; *Visual analytics.*

KEYWORDS

Explainable AI, XAI, explainability, interpretability, interactive model exploration, sequential decision-making.

ACM Reference Format:
Anindya Das Antar, Somayeh Molaei, Yan-Ying Chen, Matthew Lee, and Nikola Banovic. 2024. VIME: Visual Interactive Model Explorer for Identifying Capabilities and Limitations of Machine Learning Models for Sequential Decision-Making. In *The 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24), October 13–16, 2024, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3654777.3676323

1 INTRODUCTION

Ensuring that Machine Learning (ML) models make correct inferences based on meaningful data relationships is necessary for their broader adoption into high-stakes decision-making scenarios (e.g., in healthcare [97, 134], child welfare [25, 195], environmental analysis [131], criminal justice [88], and public safety [104]). Model engineers who train and develop ML models are the first in line to investigate their models' limitations. For example, before deployment, they need to ensure that the models make accurate decisions and capture relevant knowledge from the data. Such evaluation is necessary for real-world deployment, where the potential for social harm is high [10, 85, 89, 107], in part due to the model making decisions on unseen data with limited human oversight [59].

Existing Explainable AI (XAI) [17] systems offer explanations in different forms [122, 179] that help ML engineers evaluate their trained models. They can use such explanations to investigate and debug model limitations at different levels of granularity (e.g., using local [146], cohort or subgroup [46, 108], and global [101] explanations). Yet, many XAI systems do not support ML evaluation for diverse data types, domains, and tasks. For example, most existing XAI systems [13, 101, 146] can provide a single explanation for each single-point estimate prediction from a cross-sectional data classification model. However, the complexity of sequential decision-making models precludes them from being reduced to any single explanation. Forcing such reduction may not align with the model engineers' mental models [128, 177] of sequential decision-making, which could lead to misleading evaluation [84].

Thus, to evaluate sequential ML models, model engineers use model-specific sequential XAI tools [115, 160, 161]. However, model- [161] and domain-specific [15] visualizations in those tools may not generalize to support diverse discriminative and generative sequential ML algorithms. Although some XAI tools [21, 102, 156] tried to adapt existing feature attribution methods to time-series data, it remains challenging for model engineers to understand the long sequences of model decisions using such explanations [1].

Existing tools and toolboxes do not enable effective decomposition of sequential model decisions. Yet, such decomposition of model decisions into "what-if" scenarios (i.e., "chunks of problem-solving know-how" [69]) is essential to allow users to develop their mental models and effectively make sophisticated high-level decisions—in our case, decisions about sequential model capabilities and limitations. Although existing sequential XAI tools [21, 115, 160, 161]) can visualize sequences, they lack support for this kind of "chunking" and comparison between chunks that could help streamline sequential model evaluation.

Here, we present Visual Interactive Model Explorer (*VIME*), a domain-, data-, and model-agnostic XAI toolbox that helps model engineers to investigate their sequential ML models in different "what-if" scenarios (Figure 1). The key design insight behind VIME is that decomposing complex sequences of ML model decisions into "what-if" scenarios helps users to form explanations about each scenario. This allows users to focus on specific evaluation tasks at different levels of granularity (e.g., local explanations for specific inputs, cohort explanations for subgroups of inputs). When combined, such partial explanations can aid in comprehensive evaluation of complex sequential ML models. The key technical contribution of VIME is its flexible interactions for grouping, filtering, and zooming into sequences that we iteratively designed and implemented on top of Sankey visualization (an existing approach for visualizing sequential decision-making [136]).

To evaluate VIME, we conducted simplified user evaluation [126] using think-aloud [78, 79] with 14 ML engineers. To test VIME across sequential models trained on different datasets from different domains, we chose two state-of-the-art sequential ML models [11, 15] trained on pooled data (i.e., data that combines both cross-sectional and time-series data) from two domains: 1) predicting aggressive driving behaviors, and 2) forecasting behaviors of people with Multiple Sclerosis (MS). To compare and contrast VIME with different existing "what-if" scenario exploration tools, we selected the Google "What-If tool" (WIT) [180], a state-of-the-art baseline toolbox. WIT combines multiple explanation tools for "what-if" scenario exploration into one system and is agnostic to both ML algorithms and deployment domains. Also, we modified and improved WIT to support sequential ML evaluation by adding functionality that resembles an existing sequential XAI tool [21].

In our study, participants used VIME and WIT to reconstruct a fixed "what-if" scenario and later chose their preferred tool for independent model exploration. Our findings showed that compared to WIT, VIME made it easier for participants to create their own rules for decomposing long sequences of ML model decisions into "what-if" scenarios that they could then compare. This helped them identify instances when the model made wrong decisions (e.g., overgeneralizing to physically impossible situations, missing input-output relationship, spurious correlations), and to investigate what caused those wrong decisions (e.g., identify scenarios with missing or underrepresented data). Moreover, 12 of the 14 participants selected VIME for the independent exploration task; making VIME their preferred sequential ML model evaluation toolbox.

Our work contributes a system design and a high-fidelity functional prototype of VIME. The key design decision that differentiates VIME from existing XAI work are interactions that allow users to set their own rules for decomposing long sequences of ML model decisions into "what-if" scenarios that they can then compare. Although our selected baseline, the Google WIT [180], enables users to define and compare their own chunks it does not necessarily mean that such chunks can support sequence comparison. Thus, our work highlights the importance of interactions that are specifically designed for effectively comparing sequences [106] to support sequential ML model evaluation. The insights gained from our iterative prototyping and evaluation of VIME will inform the development of future interactive XAI mechanisms for evaluating sequential decision support systems.

2 RELATED WORK

Here, we start with a review of tools that support the ML engineering pipeline and then narrow down to specific design opportunities for improving existing XAI methods to assist ML engineers in identifying model capabilities and limitations. We distinguish methods and tools for data pre-processing, model development, and evaluation. We focus on eXplainable AI (XAI) systems that aid in model evaluation by explaining ML decisions at different levels of granularity, with a specific focus on tools that support cohort (i.e., subgroup) analysis using "what-if" scenarios for sequential models.

2.1 Data Exploration and Preprocessing

Data availability and quality significantly impact the performance of trained ML models [31, 74, 143]. Hence, before model training, users need to ensure that the data they use are properly sampled [114], correctly labeled [66], not missing relevant information [129], and free from harmful historical biases [10]. Data exploration tools [136, 137] help users analyze their data and gain insights about features and labeled outcomes [18, 19]. Users can investigate cross-sectional data [112] at a given time, time-series data [70, 123, 137, 138] collected over time, or pooled data [61] combining both. However, such tools do not immediately apply to ML model evaluation and are, thus, beyond the scope of this paper.

2.2 Interactive Machine Learning (IML)

Interactive Machine Learning (IML) [50] aids users with training [43, 45, 49, 110, 121], optimizing [9, 52, 68, 94, 135], and selecting [116, 185] their ML models during the model development stage. Such tools adopt the Human-in-the-loop (HITL) paradigm [7, 53, 193] to elicit and incorporate user feedback for iterative model adjustments and improvements. Existing IML tools [50, 65] primarily rely on quantitative performance metrics and error analysis to help users evaluate their model predictions against ground truth labels. However, the effectiveness of such analysis relies heavily on the quality of existing training data [41, 141] without providing explanations that aid users in understanding ML model decisions.

2.3 Explainable AI (XAI) Tools and Toolboxes

Existing explanation tools help users identify ML model capabilities and limitations [62]. Although some models are inherently interpretable to ML engineers [29, 54, 145], most "black-box" models require post-hoc explanations after it is trained [22, 46, 95, 176]. Existing XAI tools offer explanations in different forms (e.g., feature attributions [17, 64, 101, 146, 179], counterfactuals [81, 122]) and at different levels of granularity (e.g., local [146] for individual inputs and outputs, cohort or subgroup [30, 34, 46, 105, 106, 108] for subsets of related inputs and corresponding outputs, and global [101] across all possible inputs and outputs). To offer multi-faceted explanations, XAI toolboxes [20, 80, 92, 180, 190] combine multiple explanation tools in a single system. Additionally, existing visual analytics systems [6, 63, 189, 192] help identify model limitations [8] by enabling interactive visual exploration [119], allowing users to intuitively understand and interrogate the models' decisions.

2.4 Explaining Sequential ML Models

Most existing XAI tools [101, 146] and toolboxes [20] focus on evaluating discriminative classification models for single-point estimate predictions on cross-sectional data. However, they may not readily support sequential ML evaluation, which requires assessing model predictions at each timestep (including at the end of the sequence). Furthermore, the complexity of sequential ML decisions makes it challenging to reduce them into a single explanation by adapting existing feature attribution methods to support time-series data [21, 102, 156]. Attempts at such forced reduction may not align with the users' mental models [128, 177] of sequential decision-making and lead to misleading evaluation [84, 103, 158]. Although data- [191] and model-specific XAI tools [12, 38, 111, 115, 160, 161, 178] help evaluate certain sequential models' internal architectures, they do not generalize across different complex sequential models.

2.5 "What-if" Exploration of ML Models

Existing "what-if" exploration methods (e.g., if-then rules [40, 194], interpretable decision trees [77], and various "what-if" visualizations [32, 130, 180]) support complex ML evaluation by focusing on specific scenarios (e.g., individual inputs [92] or subgroup of inputs [181, 196] and their corresponding outputs). However, it is challenging for users to effectively decompose and chunk long sequences of model decisions with many branches into simple and manageable "what-if" scenarios using those tools. Supporting scenario-based "chunking" [69] and comparing those chunks could help users streamline sequential model evaluation. Also, existing "what-if" tools that evaluate the quality of time-series data samples [39, 58, 190] and predictions from sequential ML models [5, 38, 173] focus on explaining specific sequential models, often deployed to specific domains. As such, they may not generalize across different data types, ML models, and deployment domains.

3 VISUAL INTERACTIVE MODEL EXPLORER

Here, we describe the design of the Visual Interactive Model Explorer (VIME) that enables users to investigate their ML models interactively. We conducted a review of existing literature to understand the current context of the use of XAI tools and toolboxes and to distill the needs of our target users—ML model engineers. We designed VIME to address ML model engineers' needs because they are the first in line to develop and evaluate ML models. They have task expertise in ML models and are often highly trained computer scientists with graphical user interface (GUI) and command expertise [60] with existing ML and data science tools. We first describe the specific needs of our target users when investigating ML models and ground our design goals (DG) within those user needs. We then describe our design process, highlighting specific design choices and trade-offs that we considered. Our design centers on "what-if" scenario-based interactive model exploration to break down complex explanations into smaller parts and simplify the evaluation of complex sequential ML models.

3.1 Model Engineer Needs & VIME Design Goals

To understand model engineers' needs and challenges, we reviewed existing literature on interactive ML debugging [35, 43, 49, 124, 133, 166], visual analytics [6, 63, 189], XAI [3, 36, 51, 73, 148, 172], and Interpretability [44]. We also incorporated insights about the current context of use from existing work [26, 27, 42, 71, 99, 140, 168, 180, 187, 188] that included formative studies and semi-structured interviews with ML engineers as part of their design. We compiled a comprehensive list of user needs and validated them through discussions with ML and AI researchers (authors and others). We then described our design goals to address each need.

Although conducting another formative study with model engineers could have confirmed existing findings about the current context of use from prior work, such a formative study was not necessary to complement an already comprehensive list of user needs that the existing work has identified. We also note that design trade-offs prevent any single design from fully meeting every design goal and addressing every user need.

3.1.1 Supporting Different Stages of Model Development and Evaluation. Model engineers need to explore their models: 1) during model training and development and 2) after the model has been trained. Although their objectives may differ at these stages, their common goal is identifying ML errors and limitations. During development, they use such insights to train [43, 49], optimize [52], and select [185] the final ML algorithm. After training and before deployment, they need to understand and explain model decisions [28, 166]. Our key design goal is to facilitate post-hoc evaluation of existing trained models. Although this goal could be relevant to comparing candidate ML models during training, how these insights modify the model is beyond the scope of our work. However, users may need to compare insights about the trained model against data, making access to the training and testing data relevant.

3.1.2 Supporting Model Evaluation Across Diverse Domains. Model engineers often train models for deployment in diverse domains to support high-stake decision-making. Each domain possesses distinct phenomena that they need to ensure that their ML model has captured [63, 73, 165]. However, developing separate ML evaluation tools for every domain with tailored visualizations and interactions is resource-intensive [186]. Furthermore, some domains lack readily available tools. Thus, our design goal is to create a domain-agnostic tool that supports users in ML model evaluation across multiple domains, where the models are or will be deployed.

3.1.3 Supporting Model Evaluation Across Diverse Data Types. Model engineers train their models on different data types [61]: 1) cross-sectional, 2) time-series, or 3) pooled data that combines both. To evaluate classification models trained on cross-sectional data, they need to investigate models' predictions for each data point or sample [63, 189]. Evaluation of sequential ML models trained on time-series data requires them to understand how feature values change at each time step and influence model predictions [148]. Pooled data requires both. They also need to distinguish time-independent and time-dependent features and outcomes to facilitate efficient analysis of such models. Thus, our design goal is to support evaluations of ML models trained on all three data types.

3.1.4 Supporting Different Model Types and Evaluation Tasks. Model engineers use different ML models (e.g., discriminative [96], generative [75]) to perform classification [117] and forecasting [11, 132] tasks. They need to debug why a discriminative model predicts a specific output given the input features [73, 93, 159]. For generative models [149, 163], they need to evaluate how and why the model generates new data. Also, while evaluating sequential models, they need to consider trends and predictions at different time steps. Thus, our design goal is to support ML model evaluation across model types and the outcome estimation tasks those models perform.

3.1.5 Supporting Model Evaluation at Different Levels of Granularity. To understand the decision-making process of black-box ML models [3] (i.e., models that are not inherently interpretable [29]), users seek explanations at different levels of granularity: 1) global [101], 2) cohort (i.e., subgroup) [108], and 3) local [146]. Global explanations provide a macroscopic overview of model decisions across all possible inputs and outputs. Cohort explanations help evaluate the model in specific subgroups, highlighting limitations that may not be apparent globally. Also, local explanations offer detailed insights into individual predictions. Thus, our design goal is to support ML model evaluation at all three levels of granularity.

3.1.6 Supporting Explanations to Match User's Mental Model. Existing research [27, 83] showed that seeking explanations about ML model decisions is a sensemaking process. Model explanations that are aligned with the mental models of their users [55, 128, 177] enable them to form comprehensive, meaningful, and accurate explanations about their models [118, 139, 174, 175]. Otherwise, model explanations can be misleading [84, 86, 113], which could result in inaccurate or inconsistent model evaluation. Thus, our design goal is to provide model engineers with interactive tools that allow them to forage accompanying evidence and justifications to validate model decisions. We also want them to form and leverage explanations that enhance their understanding of the model.

3.1.7 Supporting Multifaceted Explanations. During evaluation, model engineers seek answers to various questions about their ML models [99] to gain insights that may help manage model complexity [153, 187]. This led to the development of XAI toolboxes [20, 92, 180] that combine multiple explanation tools in a comprehensive system, each addressing a distinct question about the model. Thus, our design goal is to interconnect explanation tools within a toolbox, allowing users to interactively choose and apply the tools for breaking down the evaluation into specific model-related queries rather than offering these tools as independent and isolated explanation sessions. Also, our design goal is to provide coherent and multifaceted partial explanations that, when collectively used, can simplify the evaluation of complex sequential models.

3.2 Low-fidelity Prototypes & Design Critiques

We used an iterative approach to design and evaluate VIME. One of our early key design insights was that allowing users to create and explore relevant "what-if" scenarios provides solutions to many design goals outlined in section 3.1. In particular, this approach could simplify the evaluation of complex sequential models by allowing users to explore their models from different perspectives [57], including different evaluation tasks at different levels of granularity. This could help users develop and update their mental models by synthesizing their scenario-specific insights and hypothesizing different explanations [4, 69, 106]. Thus, we focused on exploring this potentially fruitful design direction from the beginning.

We explored different ways to present the "what-if" scenarios through low-fidelity prototypes (see Appendix) of state-of-the-art static visualizations: 1) probability plots [183], 2) waffle plots [87], 3) feature attribution plots [101, 146], and 4) Sankey diagrams [147]. We illustrated these initial prototypes on two existing sequential models: 1) aggressive driving [15], and 2) Multiple Sclerosis (MS) [11]. To investigate these prototypes against our design goals, we performed design critiques that included the authors of this paper, other Human-Computer Interaction (HCI) and AI researchers, and an MS domain expert. The first author discussed the design goals with the team for further validation. Then, the team used these four prototypes to explore different "what-if scenarios" from the MS and driving model and recorded the scores and feedback. Our evaluation criteria were usability and how well the design addresses the user needs and design goals that we highlighted.

We selected the interactive Sankey visualization as our final design since it scored the highest in our design critique and emerged as the most promising feature of our designs. The waffle plots got the second-highest score. We found Sankey diagrams helpful in intuitively visualizing and understanding complex relationships and feature dependencies to understand the sequential ML model's behavior over time and in different "what-if" scenarios. This aided us in identifying trends and patterns that might be missed with simpler visualizations. Thus, Sankey diagrams not only fulfilled our need for clarity and comprehensiveness in feature representation but also significantly enhanced the interpretability of our time-series analysis, directly aligning with our design goals.

Feedback from our design critiques highlighted that static probabilistic visualizations can be challenging to understand [125] when assessing a series of decisions from complex sequential ML models with numerous feature value combinations, potentially leading to erroneous decision-making [37, 100, 142]. To address this challenge, we improved our initial low-fidelity Sankey prototypes with interactive controls based on the feedback and developed a high-fidelity functional prototype. To reinforce our design choice and validate our prototype's usability before the main user evaluation, we conducted a pilot study with four ML and HCI researchers. They explored the Sankey-based high-fidelity prototype and confirmed the effectiveness of Sankey diagrams in revealing complex trends in sequential model behavior, particularly in "what-if" scenarios. Such a pilot study also helped us refine our final designs and solve usability issues before our main user evaluation.

3.3 High-fidelity Functional Prototype

Here, we provide a detailed system description of our final VIME design and the specific design choices we have made to address model engineers' needs. To illustrate VIME, we use a running example of a hypothetical ML model engineer, named *Samira*, who wants to explore and evaluate an existing generative sequential ML model [14, 15] trained on existing time-series data [72] that can: 1) classify (i.e., discriminative predictions) if an instance of driving through an intersection is aggressive or not, and 2) generate new driving instances (e.g., to show aggressive drivers alternative, non-aggressive ways of driving). Such exploration involves interactively investigating various "what-if" scenarios to provide evidence and justifications (i.e., explanations) for why a model made a particular decision. Since that is a large sequential ML model, it would be challenging to evaluate it by reducing it to any single explanation.

Instead, VIME allows Samira to simplify model evaluation by breaking it into smaller evaluation tasks and focusing separately on each simple and relevant "what-if" scenario. We describe how Samira would start her exploration of model decisions by creating a simple scenario indicative of aggressive driving behavior that she is familiar with (Fig. 1). After completing this task, Samira would create and explore other similarly-sized scenarios to make sense of the large model by combining insights from each of those scenarios. We describe VIME features in order that Samira would use them to create her first "what-if" scenario.

3.3.1 *Internal Data and ML Model Representation.*

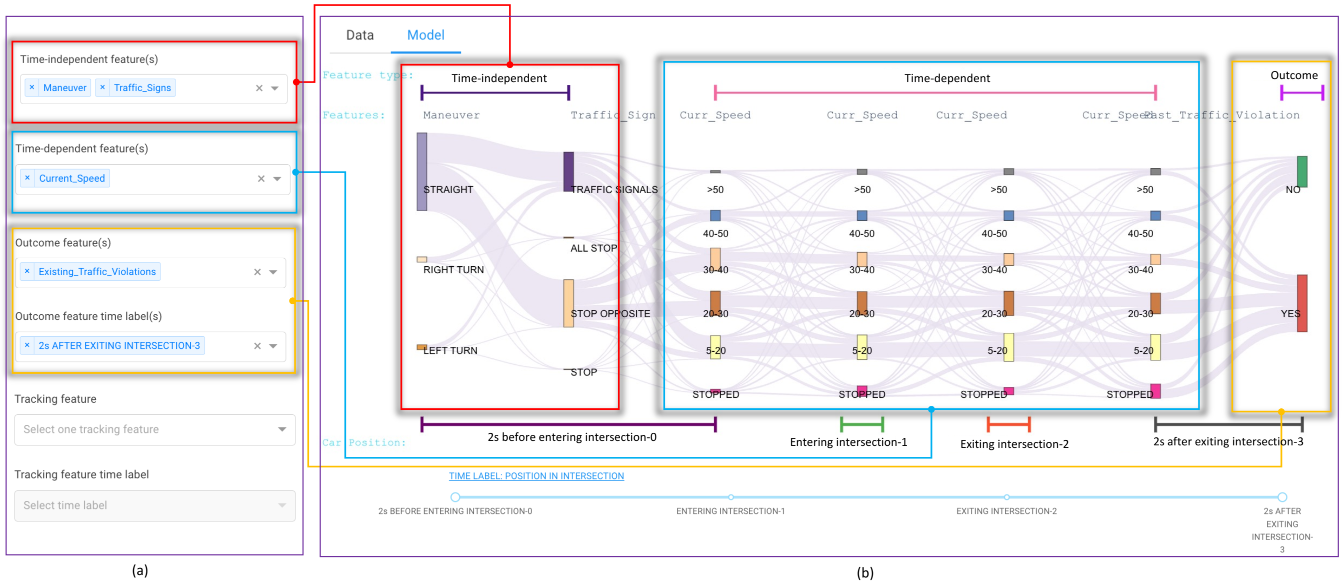
VIME maintains an internal representation of model $\mathcal{M}$ trained on data $\mathcal{D}$ to support the evaluation of different data and model types. $\mathcal{M}$ is a tuple:

$$\mathcal{M} = (\mathcal{D}, \mathcal{S}, P(s_0), P(s' \mid s)) \tag{1}$$

Here, $\mathcal{S}$ represents a set of states, where each state $s \in \mathcal{S}$ uniquely identifies possible combinations of time-independent and time-dependent feature values, represented by the feature vector $\mathcal{F}_{\mathcal{S}}$. $P(s_0)$ represents the probability that a state $s_0 \in \mathcal{S}$ initiates a decision sequence, and $P(s' \mid s)$ represents the probability of the model transitioning from state s into the next state s' in the decision sequence. VIME can then sample sequences of model decisions using the probability functions $P(s_0)$ and $P(s' \mid s)$ or display sequences from $\mathcal{D}$ independent of the model and data types.

For probabilistic sequential ML models like Markov Decision Process (MDP) [144] with time-series data, VIME directly constructs $\mathcal{M}$ from the model parameters. When data includes time-independent features, $P(s' \mid s)$ becomes deterministic. In generative models, VIME uses $P(s_0)$ and $P(s' \mid s)$ for different states in $\mathcal{S}$ to generate new samples. VIME estimates $\mathcal{S}$ from the union of all possible input and output states, with each state terminating the quasi-sequence for discriminative models with cross-sectional data. VIME then estimates $P(s_0)$ from the training data and $P(s' \mid s)$ from the quasi-probabilities of model predictions given inputs. For discriminative models with time-series data, the representation remains similar, but not every state terminates the sequence.

VIME primarily samples outputs from live models, with an option to load model outputs from a *CSV* file when the live model is not available. In our illustration, Samira has access to an existing live model [15]. She can then categorize features based on time dependency and choose the order of feature values for visualization.



Time-independent feature(s): Maneuver, Traffic_Signs

Time-dependent feature(s): Current_Speed

Outcome feature(s): Existing_Traffic_Violations

Outcome feature time label(s): 2s AFTER EXITING INTERSECTION-3

Tracking feature: Select one tracking feature

Tracking feature time label: Select time label

Figure 2: VIME's a) feature selection panel to select a subset of relevant model features, and b) interactive Sankey visualization showing a sequence of decision steps for the time-independent, time-dependent, and outcome features in order of selection.

3.3.2 *Feature Selection Panel.* It is difficult to visualize all possible inputs and outputs from a large sequential model together at the same time. Therefore, VIME's "feature selection panel" (Fig. 2a) categorizes features into three types: 1) *Time-independent Features* with deterministic transitions and consistent values, 2) *Time-dependent Features* with stochastic transitions in values that vary over time, and 3) *Outcome Feature(s)*, showing model predicted outcome(s) given one or more input features, with "outcome feature time label(s)" representing timestep(s) of the outcome feature.

To explore the large driving model, Samira starts exploration with a simple scenario. She wants to observe the "Maneuvers" (e.g., straight, left turn, and right turn) of how people are driving through an intersection with different "Traffic Signs" (e.g., stop, stop opposite, all stop, and traffic signal). Also, she is interested in how fast people are driving (i.e., "Current Speed" at four intersection positions) and whether speeding behaviors influence the model's prediction of aggressive and non-aggressive driving behavior. Thus, she selects "Existing Traffic Violations" as the outcome feature.

3.3.3 *Interactive Sequence Visualizations.* VIME leverages Sankey diagrams [147, 150] to create two distinct interactive sequence visualizations for displaying data and model samples for the selected time-independent, time-dependent, and outcome features (Fig. 2b). Users can intuitively visualize the transitions over time and complex input-output relationships to understand the dynamic nature of sequential ML models. This is particularly helpful for analyzing the influence of time-dependent features on model predictions at different timesteps, facilitating time-series feature evaluation. In these diagrams, each rectangular node corresponds to a feature value at a specific timestep, and edges represent the proportion of behavior instances between source and target nodes.

From the model tab visualization (Fig. 2b), Samira observed that drivers more frequently continued straight at intersections than turning, especially at traffic signals or stop-opposite signs indicating right-of-way. She speculated this might be because the training data had more examples of such "Maneuver" and "Traffic sign" values.

3.3.4 *Range Slider to Zoom into Specific Timestep of the Sequence.* It is challenging to evaluate long sequences of model decisions for all possible timesteps, which may complicate the visualization. To streamline the visualization, VIME offers a "range slider" to zoom in [154] and visualize model decisions at a specific timestep range. Users can drag the range slider in both directions to break down the evaluation of a long sequence and simplify their exploration.

Samira is familiar with a scenario in which aggressive drivers are likely to stop in the middle of the intersection and block it before exiting it. She wants to check whether the model captured that "speeding" while entering the intersection leads drivers to suddenly stop before they exit. Thus, she uses the range slider to zoom into car speeds from "entering" to "exiting" the intersection (Fig. 3).

3.3.5 *Interactive Filtering of Feature Values.* VIME provides "interactive filtering" of feature values to query the model with domain-relevant "what-if" scenarios. This helps users evaluate specific parts of a sequence, such as local and cohort instances, where different features take on specific values. Users need to *click* on a Sankey node to highlight the sequences with the chosen feature value, while a *shift+click* discards the value from the visualization. The selected feature values for the filter appear in the "filter values" menu, and VIME recalculates the probability distributions for the remaining feature values to ensure consistent explanations. Model engineers can then evaluate model outcomes to ensure correctness and their ability to capture meaningful data relations in these scenarios.

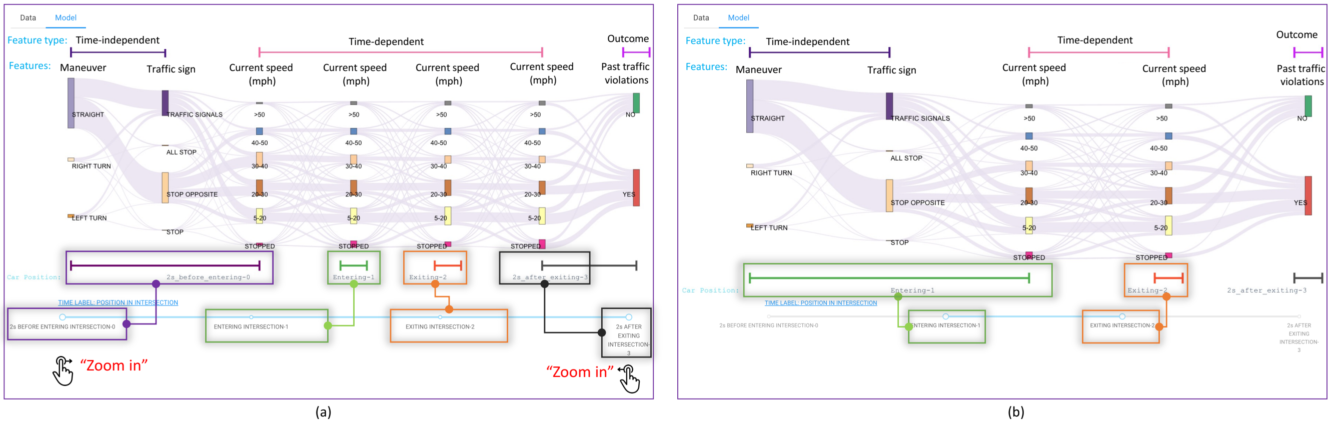


Figure 3: A range slider for zooming in and out of specific sequence timesteps showing: a) before and b) after zooming in to visualize how fast drivers were driving while entering and exiting the intersection.

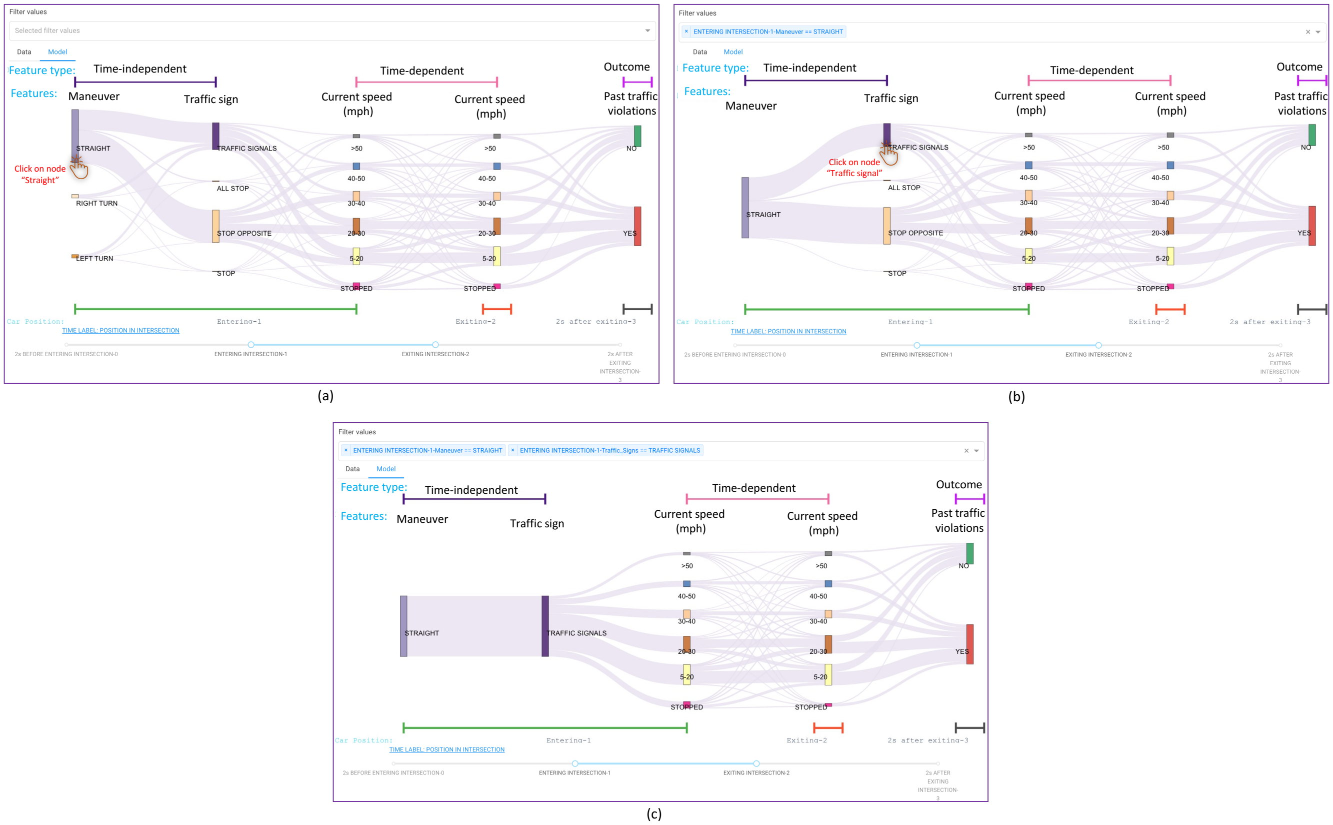
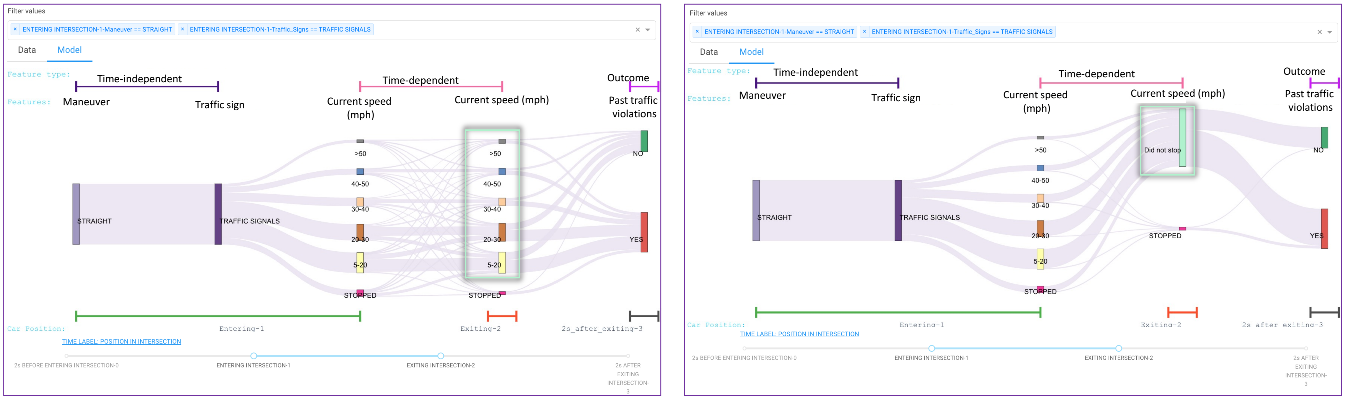


Figure 4: Interactive filtering of feature values to create a "what-if" scenario when drivers went straight through traffic signals: a) before filtering, b) after filtering "Maneuver" is "Straight", and c) after further filtering "Traffic sign" is "Traffic signals".

Samira hypothesizes that aggressive drivers going through an intersection when the traffic signal light turns yellow (i.e., "running a yellow light") may be forced to stop in the intersection, blocking it before exiting the intersection. Samira wants to identify specific instances where drivers go straight through the intersection because it is common for drivers to stop at the intersection when they are turning left or right. Therefore, she filters instances with straight maneuvers (Fig. 4a). Also, she wants to concentrate on intersections with traffic signals, so she filters only such intersections (Fig. 4b). The resulting visualization (Fig. 4c) shows only sequences of driving behaviors where the driver is going straight through intersections with traffic lights.



(a) (b)

Figure 5: Grouping feature values to reduce visualization complexity and compare behavior instances (e.g., drivers who "stopped" and "did not stop" at the exiting intersection): a) before and b) after clustering current speed values at the exiting intersection.

3.3.6 Group Feature Values. Model engineers may need to group multiple feature values with similar characteristics to reduce visualization complexity or create "what-if" scenarios. VIME enables the user to cluster multiple feature values using the "group/ungroup feature values" (see Appendix). Also, VIME recalculates the probability distribution for grouped nodes to ensure explanation consistency.

Samira wants to compare behavior instances where the cars stopped vs. did not stop while exiting the intersection. Thus, she uses the "group/ungroup feature value" interaction to create a cluster of all current speed values when the drivers are exiting the intersection and names the cluster "Did not stop" (Fig. 5).

3.3.7 Tracking Feature. Users can select a "tracking feature" to understand its relations with selected input features and the outcome. For time-dependent features, they need to specify the time step using the "tracking feature time label". The tracking feature maps its values to the edges of the Sankey diagram based on the joint probability distribution of the tracking feature and the outcome feature given the time-independent and time-dependent features. It then sets the color based on each unique tracking feature value. Contrasting color edges represent the flow of how the selected tracking feature values change in relation to other feature nodes and influence model decisions at each timestep.

Samira selects the current speed at exiting the intersection as the tracking feature (Fig. 6). VIME then updates the visualization by coloring the edges based on selected tracking feature "exiting current speed" values (e.g., stopped vs. did not stop) and shows the distribution with other features throughout the Sankey graph.

3.3.8 Selection of Distinguishable Colors. VIME allows users to select colors for feature value nodes and edges in the Sankey diagram using the "Recolor" button. Users can assign color palettes and schemes from "Color Brewer" [24, 67] based on the nature of the feature (e.g., sequential, diverging, or qualitative) and the number of feature values to be visually distinguishable [67, 157]. If a feature exceeds the palette's color limit, VIME alerts the user and suggests feature value grouping to prevent color reuse and reduce visual complexity in distinguishing features [67, 157]. Samira uses visually distinguishable colors that are colorblind safe to visualize feature value nodes and edges efficiently.

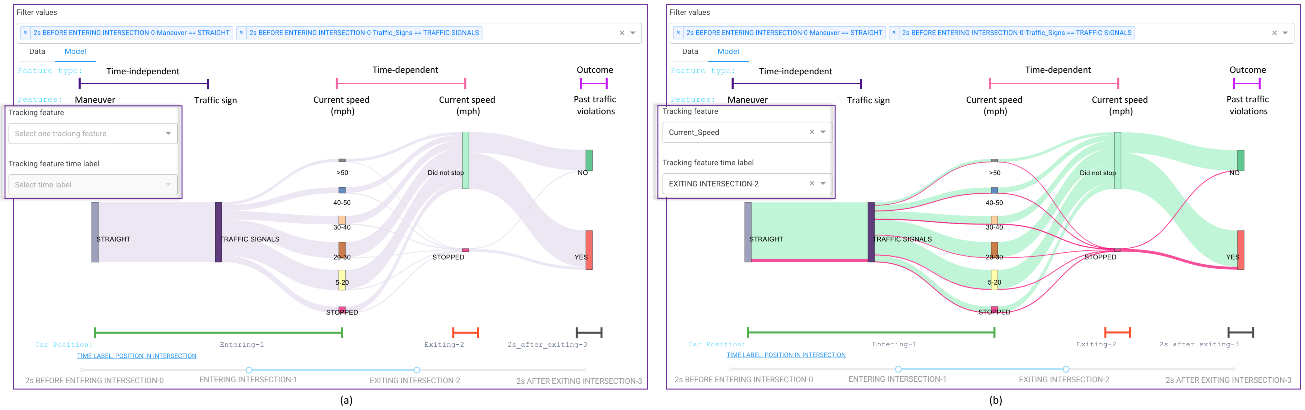
3.3.9 Hover on Nodes and Edges for Details on Demand. Offering details only when necessary [154] prevents overwhelming model engineers and helps maintain a clear and uncluttered visual space during interactive ML model exploration. The "hover on nodes and edges" functionality enables access to details during model evaluation as required. Hovering over nodes reveals the probability distribution for that feature value, while hovering over edges displays the distribution between adjacent feature values (i.e., source and target nodes), allowing users to examine their relationships.

To investigate the likelihood of drivers stopping before exiting the intersection, Samira hovers over "stopped" node (Fig. 7a), which shows a 5.1% chance that drivers will stop at "exiting intersection" according to the model. However, hovering over the edge (Fig. 7b), she observes that there is a 92.68% chance that such instances are coming from aggressive drivers with past traffic violations.

3.3.10 Comparison Between Data and Model. VIME displays Sankey visualizations for the data and the trained model in separate tabs. Any user action, such as zooming into a timestep or filtering feature values in one tab automatically reflects in the other. This synchronization ensures consistent explanations and lets users compare model-driven behaviors with real-world data scenarios seamlessly.

While evaluating the model, Samira spots an anomaly: the sequential model predicts a 2.9% likelihood that cars decelerate from over 50 mph to a full stop within 2 seconds, which is physically impossible (Fig. 8b). She then filters sequences where cars stopped at the exiting intersection and switches to the data tab, which displays that this scenario is underrepresented in the data (Fig. 8a).

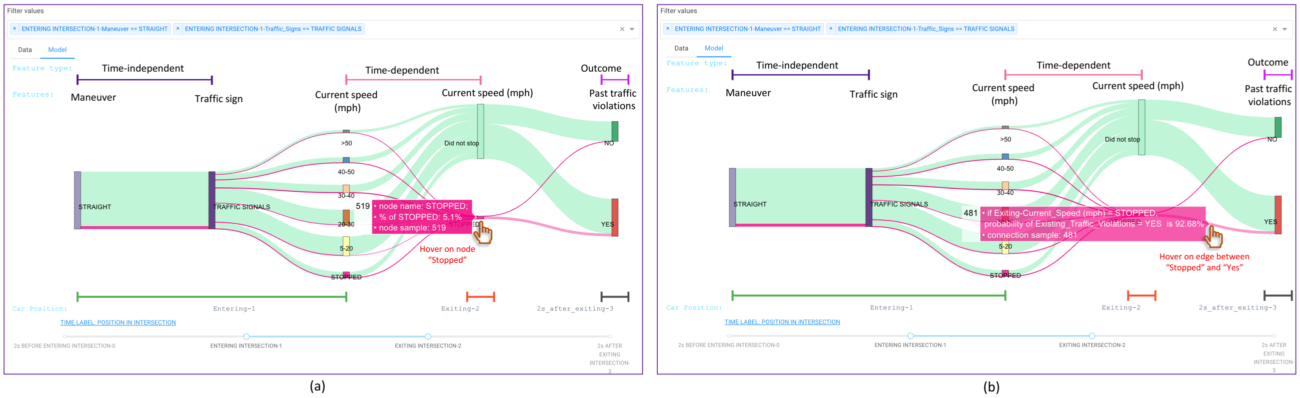
3.3.11 Live Sampling and Sample Size Determination. Limited sample sizes in the visualization can restrict the model engineers from drawing meaningful conclusions about model predictions. VIME provides a "sample size determination panel", displaying quantitative summary statistics, such as the total sample size, filtered sample size, confidence interval ($1 - \delta$), and absolute error (ϵ) for


Filter values: 2s BEFORE ENTERING INTERSECTION-0-Maneuver == STRAIGHT; 2s BEFORE ENTERING INTERSECTION-0-Traffic_Signs == TRAFFIC SIGNALS

(a) Tracking feature: Select one tracking feature; Tracking feature time label: Select time label

(b) Tracking feature: Current_Speed; Tracking feature time label: EXITING INTERSECTION-2

Figure 6: Tracking feature sets the edge colors and displays the probability distributions in relation to other feature values: a) before and b) after applying the “Current speed” before exiting the intersection as a tracking feature.


Filter values: ENTERING INTERSECTION-1-Maneuver == STRAIGHT; ENTERING INTERSECTION-1-Traffic_Signs == TRAFFIC SIGNALS

(a) 519
- node name: STOPPED;
- % of STOPPED: 5.1%
- node sample: 519

(b) 481
- if Exiting-Current_Speed (mph) = STOPPED, probability of Existing_Traffic_Violations = YES is 92.68%
- connection sample: 481

Figure 7: Hovering on nodes and edges displays detailed information; in this scenario, the likelihood of: a) drivers stopping at the exiting intersection, and b) aggressive driving instances (i.e., drivers had traffic violations) when they stop at exiting.

data and model samples (Fig. 8c). We estimated ϵ under fixed confidence $1 - \delta = 80\%$ per. This method considers the total feature value combinations and available filtered samples to estimate the absolute error. If the error exceeds the default acceptable threshold ($\epsilon < 0.05$ under $1 - \delta = 80\%$) [91], VIME alerts users with a “caution message”, guiding them to interpret model outcomes carefully.

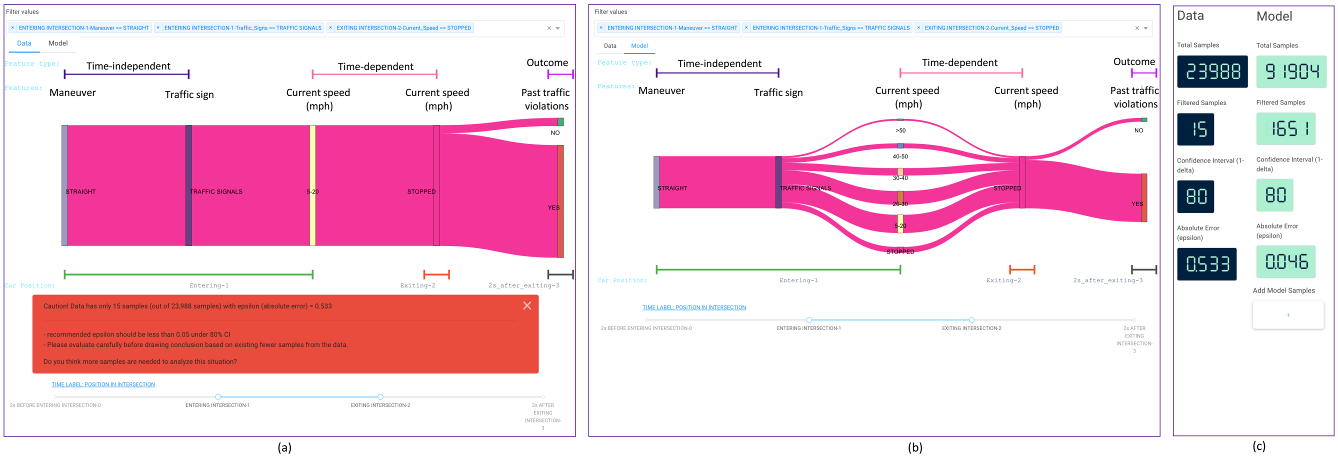
The “add model samples” button (Fig. 8c) enables live sampling from the model $\mathcal{M}$ using the probability distributions in Equation 1. For each sequence, our algorithm first samples an initial state s_0 from the distribution $P(s_0)$. We then sample the next state s_i in the sequence using the transition probabilities $P(s_i|s_{i-1})$. This continues until we encounter an end state that terminates the sequence. If only CSV files are included, this button remains disabled.

Samira notices a “caution message” (Fig. 8a), indicating that instances of stopping at intersections are underrepresented in the data with just 15 samples (Fig. 8c), where the ϵ exceeds the acceptable threshold of 0.05. She concludes that the model (Fig. 8b) may overgeneralize to unrealistic speed transitions due to missing data. Samira concludes that the model could benefit from introducing knowledge about the physics of the vehicle movement.

After completing this task, Samira can now explore other relevant intersection features (e.g., intersection layout, speed limit) related to the current “what-if” scenario or explore new similarly-sized scenarios of her interest (e.g., accelerating or decelerating rapidly at stop signs). She can then combine insights from each of these scenarios to make sense of the large model across scenarios.

3.3.12 Domain-agnostic Visualizations and Interactions. Model engineers create ML models trained using various algorithms with data from diverse domains. VIME visualizations and interactions are model- and domain-agnostic. For example, just as Samira used VIME to explore the driving model for classifying aggressive driving, she can use it to investigate another sequential ML model [11] that forecasts the physical functioning of people with MS.

To evaluate the MS model, Samira uses VIME to create a “what-if” scenario where people with MS start their day with high fatigue, suspecting it correlates with low functioning. She selects age as time-independent, fatigue as time-dependent, and end-of-day (EOD) lower extremity functioning (LEF) as the outcome. She then zooms into the wake and morning time intervals on the range slider and



Filter values: ENTERING INTERSECTION-1-Maneuver == STRAIGHT; ENTERING INTERSECTION-1-Traffic_Signs == TRAFFIC SIGNALS; EXITING INTERSECTION-2-Current_Speed == STOPPED

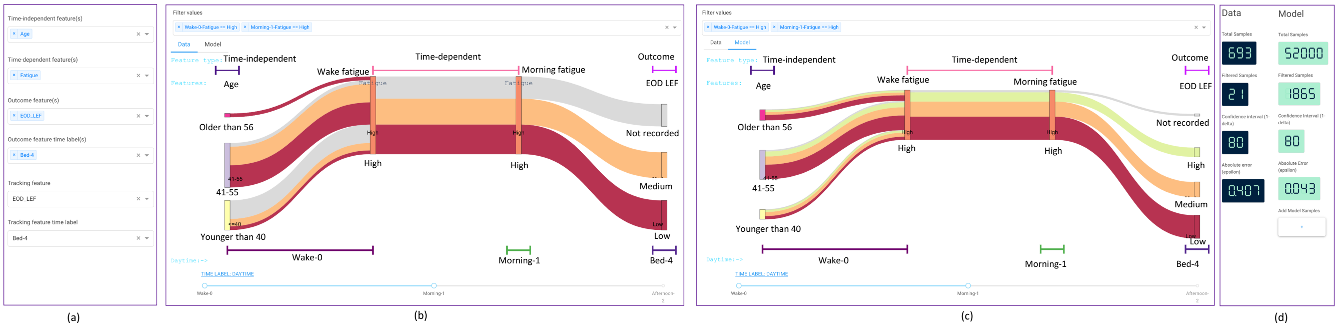
Caution! Data has only 15 samples (out of 23,988 samples) with epsilon (absolute error) = 0.533

- recommended epsilon should be less than 0.05 under 80% CI
- Please evaluate carefully before drawing conclusion based on existing fewer samples from the data.

Do you think more samples are needed to analyze this situation?

	Data	Model
Total Samples	23988	91904
Filtered Samples	15	1651
Confidence Interval (1-delta)	80	80
Absolute Error (epsilon)	0.533	0.046

Figure 8: What-if scenario from the driving model [15], where drivers stopped just before exiting the intersection: a) data visualization tab, b) model visualization tab, and c) sample size determination panel. This scenario is under-represented in the data, causing the model to over-generalize to physically impossible situations (e.g., entering the intersection at 50 mph and decelerating to a full stop before exiting the intersection).



Time-independent feature(s): Age

Time-dependent feature(s): Fatigue

Outcome feature(s): EOD_LEF

Outcome feature time label(s): Bed-4

Tracking feature: EOD_LEF

Tracking feature time label: Bed-4

Filter values: Wake-0-Fatigue == High; Morning-1-Fatigue == High

	Data	Model
Total Samples	693	52000
Filtered Samples	21	1865
Confidence Interval (1-delta)	80	80
Absolute error (epsilon)	0.407	0.043

Figure 9: What-if scenario from the MS model [11], where people with MS start their day with high fatigue. a) Feature selection panel, b) data visualization tab, c) model visualization tab, and d) sample size determination panel. The model has generalized to situations that are missing in the data (e.g., recovery towards high end-of-day (EOD) Lower Extremity Functioning (LEF)).

filters for high fatigue instances. Samira notices that this "what-if" scenario is underrepresented in the data, with just 15 records, and shows only "medium" or "low" EOD LEF (Fig. 9a). However, the model correctly generalized to unseen data to forecast "high" EOD LEF for people who are less fatigable [155]—people with MS that can continue to perform physical activity despite feeling physical fatigue (Fig. 9b). Thus, Samira concludes that although overgeneralization can lead to incorrect outputs from the model, correct generalization can also yield accurate predictions from the model.

4 USER EVALUATION

We conducted a simplified user evaluation [126] with 14 ML engineers using the think-aloud method [78, 79] to assess VIME. To scope our evaluation, we focused on investigating trained ML models before their deployment for real-world sequential decision-making. Thus, we selected two off-the-shelf sequential ML models [11, 15] trained on existing datasets [72, 90] containing pooled data from two domains. Our main goal was to compare VIME with different existing tools for "what-if" scenario exploration. Thus, we used Google What-If Tool (WIT) [180], which combines state-of-the-art "what-if" scenario exploration tools into one system. To facilitate sequential ML evaluation, we modified WIT by loading it with data and predictions for each timestep and adding features like timestep-specific Shapley values (similar to TimeSHAP [21]).

4.1 Datasets and Models

We chose two existing sequential ML models (Table 1) from two domains to: 1) classify aggressive driving behaviors [15], and 2) forecast end-of-day physical functioning of people with Multiple Sclerosis (MS) [11]. Both models use a Markov Decision Process (MDP) for sequential decision making, along with different ML algorithms to estimate initial state probabilities $P(s_0)$ and transition probabilities $P(s'|s, a)$, and an Inverse Reinforcement Learning (IRL) algorithm to estimate action probabilities $P(a|s)$. Those underlying probability distributions map onto VIME's internal model representation, but without loss of generality (Section 3.3.1).

Table 1: Comparison between MS and Driving sequential ML models and datasets.

Comparison factors	MS dataset [90] and model [11]	Driving dataset [72] and model [15]
Total sequences in dataset	693 behavior sequences of 107 people with clinically-identified Multiple Sclerosis (MS)	23,988 behavior sequences of 26 licensed drivers as they daily drove through intersections of a mid-sized city in North America
Type of data	Pooled data (both cross-sectional and time-series)	Pooled data (both cross-sectional and time-series)
Type of model	Generative sequential model	Generative sequential model
Type of evaluation task	Forecasting end-of-day (EOD) functioning of people with MS	Classify aggressive driving instances
Type of ML algorithm	Bayesian Network with Inverse Reinforcement Learning	Bayesian Network with Inverse Reinforcement Learning
Each sequence length and timestep	Five daytime intervals (wake, morning, afternoon, evening, and bed)	Four positions in the intersection (2s before entering, entering, exiting, and 2s after exiting)
Ground truth(s) and model outcome	Self-reported levels of (low, medium, high) EOD: 1. Lower extremity functioning (LEF) 2. Upper extremity functioning (UEF)	Drivers had past records of existing traffic violation (Yes, No)
State features	1. **Timestep:** daytime intervals 2. **Time-independent features:** 2.1. **Demographics:** Gender, Age 2.2. **Health condition:** MS subtype, Mobility aids 3. **Time-dependent features:** 3.1. **Symptoms (self-reported):** Pain, Fatigue 3.2. **End-of-day functioning (survey):** EOD LEF, UEF	1. **Timestep:** car position in the intersection 2. **Time-independent features:** 2.1. **Environment:** Intersection layout, Traffic signs, Maximum speed limit, Rush hour 2.2. **Destination goal:** Maneuver 2.3. **Driving record:** Existing traffic violations 3. **Time-dependent feature:** 3.1. **Vehicle current state:** Current speed
Action features	**Time-dependent features:** 1. Activity bouts (activity intensity), 2. Activity pace (slowing down in between activities)	**Time-dependent feature:** Pedal (order of pressing gas or/and break, such as, breaking soft, throttle soft, braking hard)
Model's outcome estimation performance	1. ROC AUC for LEF forecast: 0.78 2. ROC AUC for UEF forecast: 0.85	≈ 85% accuracy in classifying aggressive driving instances

Table 2: User study participants' demographics and expertise who evaluated MS or driving data and model.

	Gender	Age	Occupation	Race/ Ethnicity	Research area	Research experience	Domain knowledge	Taken ML class before	Implemented/used ML models	Used XAI systems
PA01	Woman	22-25	PhD student	Asian	Computer science, ML/AI	3+ years	No	Yes	Yes	Yes
PA02	Man	26-30	PhD student	Asian	Computer science, Data science, ML/AI	4+ years	No	Yes	Yes	Yes
PA03	Man	31-35	PhD student	Asian	Computer science, Data science	5 years	No	Yes	Yes	Yes
PA04	Woman	22-25	Masters student	Asian	Computer science, HCI	2+ years	No	Yes	Yes	No
PA05	Woman	22-25	Masters student	Asian	ML, HCI	2+ years	No	Yes	Yes	Yes
PA06	Man	26-30	PhD student	White	Computer science, HCI, ML/AI	6 years	No	Yes	Yes	Yes
PA07	Woman	22-25	PhD student	White	Computer science, ML/AI	2+ years	No	Yes	Yes	Yes
PA08	Man	22-25	Masters student	Asian	Computer science, ML/AI	1+ years	No	Yes	Yes	Yes
PA09	Woman	22-25	PhD student	Asian	Data science	3+ years	No	Yes	Yes	Yes
PA10	Woman	26-30	Masters student	Asian	ML, HCI	2 years	No	Yes	Yes	Yes
PA11	Man	22-25	Masters student	White	Computer science, ML/AI	2+ years	No	Yes	Yes	Yes
PA12	Woman	22-25	Masters student	Asian	Computer science, ML/AI	2 years	No	Yes	Yes	No
PA13	Man	22-25	Masters student	White	Computer science, Data science	2 years	No	Yes	Yes	Yes
PA14	Man	18-21	Masters student	White	Data science, ML/AI	3 years	No	Yes	Yes	No

4.2 Choice of Baseline Toolbox for Evaluation

Existing interactive visualizations (e.g., waffle [87], Sankey [147], probability plots [183]) and explanation tools (e.g., partial dependence plots (PDP) [23], counterfactuals [171], Shapley values [21]) support "what-if" exploration in different forms. We selected the Google What-If Tool (WIT) [180] because this toolbox combines most of those explanation tools for "what-if" exploration at different levels of granularity. For example, WIT includes Facet Dive [130] tool, which displays model inputs and outputs in waffle charts for exploring local and subgroup-level "what-if" scenarios, similar to visualizations that we explored in Section 3.2. WIT users can hypothesize and validate various explanations for the effects of features on model outcomes by consulting various summary statistics, identifying counterfactuals [171], and visualizing PDP [23].

Existing surveys [82, 98] and user studies demonstrated the applicability of WIT in evaluating ML models trained on cross-sectional [109, 181] and time-series [2, 56, 152] data. Following those studies, we loaded the WIT tool with data and predictions (using the WIT "custom prediction function" API) for each timestep of the sequence to support both time-independent and time-dependent inputs and outputs. Also, we further modified and improved WIT (using the WIT "feature attribution" API) by adding support for timestep-specific Shapley values, similar to TimeSHAP [21].

Although other XAI tools [21, 102, 156] may support some aspects of sequential model exploration, they lack many of WIT's features or its broad applicability across different domains, data types, and models. For example, Patient2vec [191] is a domain- and data-agnostic tool tailored to evaluate sequential models trained

Table 3: Closed coding scheme based on the user needs and design goals that we have identified in Section 3.1.

Closed coding categories	Codes
Supporting diverse deployment domains	• Domain-specific lived experience • Data and ML-centric expertise
Supporting diverse data features	• Influence of time-independent features • Influence of time-dependent features
Supporting different models and evaluation tasks	• Evaluate discriminative properties (e.g., classify, forecast) • Evaluate generative properties (e.g., generate a sequence)
Supporting different levels of granularity	• Broad overview through global exploration • Subgroup-level insight through cohort exploration • Validate individual decisions through local exploration
Supporting user's mental model	• Ability to forage comprehensive evidence • Meaningfulness and understandability of explanations • Validate accuracy of explanations and justifications • Effectiveness of explanations to highlight ML limitations
Supporting multi-faceted explanations	• Continuity between explanation tools • Synthesize insights across diverse "what-if" scenarios

on Electronic Health Record (EHR) data. LSTMVis [161], Seq2SeqVis [160], and ProtoSteer [115] focus on internal architecture evaluation of specific sequential models (e.g., RNN-LSTM). Even though the standalone TimeSHAP [21] tool visualizes time-series Shapley values, it cannot create "what-if" scenarios.

4.3 Participants

We recruited participants from a Computer Science and Engineering graduate student mailing list who were 18 years or older and who had prior experience in implementing or using ML models (Table 2). To ensure their ML expertise, participants completed a screening survey, which asked them if they had taken ML courses and if they had practical experience in developing, using, and evaluating ML models. Note that although it may be convenient to recruit this sample, our participants were graduate students with expertise comparable to current ML engineers in the industry. We collected simplified user testing data until we observed data saturation [127], stopping at 14 participants. The participants were between 18 and 35 years old (7 men and 7 women). We compensated participants by mailing checks ($15 per hour for up to 2 hours).

4.4 Study Design

We conducted in-person simplified user testing [126] with think-aloud [78, 79]. Since no participant had specific domain knowledge, we randomly assigned each new participant to evaluate either the MS or the driving model until data saturation, stopping at eight participants for the MS model and six for the driving model. We compared VIME and WIT under identical conditions and tasks to assess their strengths and weaknesses for addressing participants' needs and challenges during sequential model evaluation.

4.5 Tasks and Procedures

After arriving at our lab, participants gave verbal consent after reading the consent form to proceed. We explained the study tasks and objectives for evaluating sequential ML models using XAI tools and showed brief video tutorials on VIME and the baseline WIT without disclosing their names. Participants then performed two tasks to evaluate either the driving or the MS model. Each session lasted approximately 2 hours. The study was reviewed and approved as exempt from ongoing oversight by our university's Institutional Review Board (IRB).

4.5.1 *Task 1: Evaluating a Prescribed "What-if" Scenario.*

Task 1 asked participants to recreate a prescribed "what-if" scenario using VIME and WIT in counterbalanced order to evaluate if it is possible to create and interpret the resulting visualizations to draw insights about model decisions. For the driving model, the scenario reflected aggressive driving behaviors identified by driving instructors [15]: "*what if drivers stop while exiting an intersection with traffic signals when going straight*" (Fig. 8). For the MS model, we consulted with a domain expert to select a relevant scenario that adversely impacts people's end-of-day functioning: "*what if people with MS begin their day feeling highly fatigued*".

4.5.2 *Task 2: Evaluating Custom "What-if" Scenarios.*

In Task 2, participants had to select their preferred tool to come up with, create, and interpret their own "what-if" scenarios. They investigated the correctness of sequential ML decisions and the ability to capture meaningful input-output relationships in those scenarios. This task also tested each system's ability to provide explanations at different levels of granularity. We also asked them to explain the reasons behind their selected tool for Task 2.

4.6 Analysis Method

Our qualitative analysis evaluated how well VIME and WIT meet the user needs and design goals outlined in Section 3.1. We transcribed think-aloud audio sessions using online tools and imported the transcripts, audio, and screen recordings into *NVivo* software. To perform closed coding [182] on the user evaluation data, we developed a codebook with initial codes falling into categories corresponding to the user need that we derived in Section 3.1.1. These categories excluded the need to support various evaluation stages, focusing instead on evaluating existing trained ML models.

Initially, we tested our codes on a subset of data to assess their applicability and made necessary refinements. The first and second authors independently conducted pilot coding in four study sessions. They discussed the pilot sessions' findings to calibrate, reach a consensus, and refine the codebook until all authors agreed. We listed the final, refined codes under each category in Table 3. The two authors then applied the codes across all 14 study sessions, periodically reviewing the data to ensure alignment with the codebook. We kept detailed memos with examples and quotes, and observations on how tool features aided participants' explorations.

4.7 Limitations

Performing quantitative user evaluation could have provided insights into the magnitude of the tool's usability and effectiveness (e.g., measuring task completion times, counting errors, and collecting self-reported usability ratings). However, our qualitative analysis still provided nuanced data necessary to identify and describe such usability issues. Also, closed coding enabled us to report relevant user interactions, insights, breakdowns, and quotes to objectively evaluate participants' preferences for WIT and VIME.

We compared VIME with only one baseline toolbox, instead of comparing it with other XAI tools, too. However, comparing standalone tools with VIME would be unfair due to their limited features compared to VIME and WIT. Also, we focused on post-hoc evaluation of trained ML models, making tools that support changes in the model during development beyond this paper's scope.

5 RESULTS

Here, we present findings from our qualitative user evaluation grounded in closed coding (Table 3). Since participants evaluated pre-trained sequential models, our findings relate to post-hoc ML evaluation after training and before deployment. We highlight key findings from the user study tasks and the consistency of participant outcomes. In Task 1, all participants identified similar types of model limitations because they had to recreate the same scenario but came up with a variety of scenarios with a diverse set of explanations in Task 2, thus showing the versatility of VIME. Our results show to what extent specific functionalities of VIME and WIT (Table 4) accomplish our design goals and meet user needs.

5.1 Supporting Diverse Domains

Both VIME and WIT were designed to generalize to data from any domain. Although neither VIME nor WIT prevented participants from investigating models from different domains, neither offered specific domain support. Thus, lacking formal expertise in the two domains, participants had to leverage their lived experiences to create different "what-if" scenarios and validate their hypotheses about the models. For example, participants who drove came up with scenarios they thought could be indicative of aggressive driving, such as *"overshooting stop signs ... while entering the intersection"* (PA06), *"[accelerating] to beat a [changing] traffic signal [yellow or red]"* (PA07), *"speeding well above the [posted] limits"* (PA03), and *"rapid acceleration followed by harsh braking"* (PA14).

Participants further interpreted the data used to train the models as the ground truth (often blindly assuming the quality and provenance of the data) and used it to justify model decisions:

> *"Model shows ... people [with MS] ... aged over 56 and using mobility aids ... likelier to have low functioning ... such reasoning makes sense ... you should ask an MS clinician to verify such outcomes further." –PA08*

Still, recognizing their lack of domain knowledge, some participants recommended cautious re-evaluation of their conclusions before deploying the models for real-world decision-making.

5.2 Supporting Different Data & Feature Types

VIME was more helpful to participants than WIT when evaluating the two models trained on pooled data. Both tools helped simplify "what-if" scenarios by selecting and filtering subsets of relevant features. However, the 2D waffle chart (WIT's primary way of visualizing relationships between features) did not allow participants to visualize the influence of more than two features at a time. Thus, participants struggled to expand long sequences and visualize branching in time-dependent feature values when using WIT. When creating the prescribed scenario in which a car stops at a traffic signal intersection proceeding straight, PA07 mentioned:

> *"While evaluating a scenario with maneuvers & traffic signs [time-independent], exiting speed [time-dependent], and traffic violations [outcome] ... comparing only two features at a time [in WIT] is burdensome." –PA14*

Adding Shapley value visualization for time-dependent features to WIT helped participants visualize those features' importance when predicting outcomes across different timesteps. For example, PA11 observed that "Pain" and "Fatigue" at bedtime have the highest Shapley values when forecasting physical functioning, but could not determine *"if morning pain and fatigue [Shapley] values have any residual impact on their high [Shapley] values at bedtime"* (PA11). Thus, helping participants interpret individual feature contributions at each timestep does not necessarily help them understand the relationship between different time-dependent features.

In contrast, VIME gave participants control over the number of features they wished to visualize and evaluate. Using VIME's interactive Sankey diagrams, they could visually track the flow and relationships of different features and their influence on the outcome, regardless of the feature type. They also used the range slider to zoom in and view those feature values within a particular range of timesteps while keeping track of the outcome features.

5.3 Supporting Different Model Types

VIME outperformed WIT for evaluating models' discriminative abilities (e.g., classifying sequences of driving behaviors as aggressive or non-aggressive), primarily due to better support for time-series data and time-dependent features, as described in Section 5.2.

VIME was also more effective in helping participants investigate the two models' generative properties. This is because WIT was primarily designed to evaluate discriminative ML models. WIT does not have a feature to automatically generate more data from the models; instead, it focuses on visualizing existing data used to train and test the models. With WIT, participants manually created new data points by editing existing individual sequences already present in the data (similar to the Prospector tool [92]):

> *"I can edit a [waffle] cell to generate a missing sample and see how it affects model predictions ... [WIT] doesn't let me generate multiple samples." –PA02*

In contrast, VIME has built-in features that support generative model evaluation. For example, PA04 used those features to generate scenarios not present in the data to explore the differences between aggressive and non-aggressive driving behaviors:

> *"The model associates sudden stops at high speeds with past traffic violations ... [VIME] lets me filter and create model samples with no violations, where cars approach intersections at the speed limit and stop gradually without hard braking" –PA04*

This example highlights VIME's utility in evaluating generated sequences and later using those sequences to classify outcomes.

5.4 Supporting Different Levels of Granularity

It was easier for participants to create global and cohort (subgroup) level "what-if" scenarios with VIME than WIT. To create global scenarios with VIME, participants simply selected relevant features without applying any filters. They then applied filters to create cohorts. They often used time-independent features to "anchor" their cohorts to model features that do not change within a sequence (e.g., specific intersection layout, specific MS subtype).

However, participants struggled to create local explanation scenarios in VIME (i.e., scenarios showing a single sequence). They could not select a specific sequence from the Sankey visualization; they had to repeatedly apply filters until they reached a single

Table 4: Functionalities of VIME and WIT influencing participants' experiences during sequential ML evaluation.

Functionality criteria	Visual Interactive Model Explorer (VIME)	Google What-If Tool (WIT)
Domain-agnostic visualization	Interactive Sankey visualization	Interactive waffle chart
Evaluating cross-sectional and time-series data features	Separate representations to visualize time-independent and time-dependent inputs and outputs	We separately load the time-independent and time-dependent features and their Shapley values
Evaluating sequential models	Internal data and model representation to support sequential ML	We modified and improved WIT to support sequential ML
Creating "what-if" scenarios at different levels of granularity	Filter, zoom-in/out, and feature value grouping to create local, sub-group, and global "what-if" scenarios	Facet dive tool to bin and edit data points to create local, sub-group, and global "what-if" scenarios
Understanding feature value distributions and relations	Details on demand about feature distributions and relationships in relevant scenarios	Time-dependent Shapley values, counterfactuals, and Partial dependence plots (PDP) to identify key feature contribution
Explanation sessions with each tool in the toolbox	Multifaceted "what-if" scenarios help combine insights from partial explanations to simplify complex ML evaluation	Each explanation tool offers independent explanation sessions to evaluate the model outcome

sequence. With WIT, participants could quickly drill down to individual sequences by selecting them in waffle plots or using the counterfactual tool to identify the "closest" counterfactual example. Thus, VIME requires support for quickly selecting and viewing local individual instances as the smallest unit of sequence comparison.

Using WIT, participants easily viewed the built-in global Shapley value plots and Partial Dependence Plots (PDP) to assess feature importance. However, they struggled to make sense if feature importance in those plots holds across different levels of granularity:

> *"... high positive Shapley values for Traffic Sign and Maneuver ... seems very important to [predict] aggressive driving ... I can't understand why they stand out ... whether the influence is scenario-specific." –PA11*

In the example above, PA11 struggled to understand if "Traffic Sign" and "Maneuver" features are also important for specific scenarios (e.g., for different intersection types or at different driving speeds). Participants could easily address such confusion in VIME:

> *"I exactly know which [what-if] scenario I am evaluating ... I can control the [level of] granularity I want for the [Sankey] visualization using filters" –PA13*

This is because they could easily select features they wanted to visualize using VIME, and modify which features (and their corresponding values) they wanted filtered in the filter menu.

5.5 Supporting User Decision Making

VIME's approach to scenario-based interactive model exploration helped participants derive more meaningful and accurate explanations compared to WIT. We attribute this to the match between VIME scenario-based sequence visualization and the participants' mental model of sequential decision-making. Also, in WIT, participants had to work to reconcile the inconsistencies across different explanation tools and their contradictory outputs even for the same "what-if" scenarios:

> *"Shapley values show high influence of MS subtype and age for forecasting functioning in females with MS ... the Partial Dependence Plot contradicts." –PA01*

After creating "what-if" scenarios, VIME interactions enabled participants to seek and obtain evidence for sequential model decision explanations they derived. They validated outcomes against the "ground truth" using the data tab, where scenarios are mirrored:

> *"The model shows cars stopping [at intersections] from over 50mph ... seems impossible ... such transition doesn't exist in [VIME's] data tab ... model overgeneralized to missing data scenarios." –PA03*

PA03 observed the driving model's tendency to overgeneralize in missing data scenarios, applying uniform probability to outcomes.

Participants also identified limitations due to missing features:

> *"Stopping at a green light intersection having the right of way could be less aggressive compared to red light or all stop intersections ... without [traffic] light color feature in the model, I can't confirm." –PA07*

Thus, participants were able to identify not only missing features but also latent domain knowledge that could have helped improve the model and their ability to evaluate model correctness.

5.6 Supporting Multi-faceted Explanations

VIME and WIT both offered a toolbox integrating various explanation tools to offer different perspectives on the data and model outputs. VIME's centralized tools and visualizations allowed for continuity between different questions participants wanted answered. However, participants noted that the tools in WIT were isolated, each providing separate explanation sessions:

> *"... challenging to re-create the same scenario when switching from Shapley values to Partial Dependence Plots. I wish the system [WIT] restored the evaluation scenario across different explanation tools." –PA09*

Thus, having to recreate the same scenario in each of WIT's diverse tools broke the sensemaking flow for the participants.

In contrast, VIME's interface maintained continuity across interactive tools to control the Sankey visualizations. This allowed participants to simplify evaluating long sequences by breaking them, step-by-step, into smaller, manageable "what-if" scenarios.

5.7 User Preferences for VIME and WIT

Out of 14 participants, 12 (85%) chose VIME for the final task, showing their preference towards VIME for evaluating sequential ML models. When asked, they attributed this choice to their initial experiences of creating a fixed "what-if" scenario with both tools in Task 1, the learning curve of each explanation tool, the usability of visualizations and interactions, the simplicity of identifying errors, and the ability to evaluate time-dependent features at different granularity. Also, among PA06 and PA07, who initially chose WIT for Task 2, PA06 switched to VIME and mentioned:

> *"I wanna compare T-type and four-way intersections ... no way I can visualize 24 possible [intersection] layouts in the waffle [chart] ... at least in the 1st tool I could group ... can I change [to VIME]?" –PA06*

Table 5: Summary of design implications for interactive explanation tools for sequential ML evaluation.

Discussion points	Breakdowns	Design implications
"What-if" exploration simplifies evaluation	"What-if" scenarios with time-dependent features over many timesteps complicates visualization	• Sliding and shifting controls in range sliders for adaptable timestep adjustments • Grouping time-dependent features across timesteps
	Challenging to recall and apply insights from previously explored "what-if" scenarios	• Interactive cards for recalling past explorations with "what-if" scenarios
Supporting foraging and sensemaking	Inadequate evidence foraging and misunderstanding explanations may lead to misleading evaluations	• Highlight parts of Sankey visualization to trace model decision pathways • Provide summary-based explanations and gather user feedback
	Without clear warning, users can develop over- or under-reliance on model decisions	• Caution message to communicate model limitations and uncertainties
Designing for broader stakeholders	Explanations tailored for computer science-savvy experts may not aid end-users	• Explanation systems catering to broader end-users (domain experts, policymakers, and consumers) needs and expertise during evaluation

The built-in features of VIME, specially designed for large sequential ML model exploration, offered a more usable alternative to the current state-of-the-art WIT for this evaluation task. Here, we credit VIME's close integration between tools that supported continuity in model exploration and aided participants in their decision-making and evaluation. This is something that other existing toolboxes do not currently include but should.

We observed no notable differences in task completion times, with each participant allocated a fixed time for Tasks 1 and 2 to identify model errors and limitations in given or chosen scenarios.

6 DISCUSSION

We contextualize findings and takeaways from our user evaluation to further improve sequential model evaluation systems. We highlight design implications [167] for user-centered interactive explanations that may simplify "what-if" scenario creation and evaluation, improve foraging and sensemaking, and cater to the diverse needs and expertise of a broad group of end-users (Table 5).

6.1 "What-if" Exploration Simplifies Evaluation

VIME's multifaceted interactions allowed users to derive "what-if" scenario-based explanations and combine them to holistically evaluate large sequential models. Instead of visualizing every feature all at once, users could simplify their analysis into familiar and manageable "what-if" scenarios. Similar to existing rule-based [69] and hypothesis-testing [106] methods, interactive visual model exploration may help users synthesize scenario-specific prior knowledge and intuition [33] during model evaluation.

Future interactions could offer users even more control to customize and simplify "what-if" scenarios. For sequences with many timesteps, sliding and shifting windows in the range slider may allow quick selections of specific timestep ranges for visualization. Grouping time-dependent features across timesteps could further simplify analysis. Also, users may struggle to recall and apply insights from previously explored scenarios to new ones. Therefore, it is important that future model exploration interfaces facilitate tracking explanations across scenarios (e.g., using interactive cards for scenario-specific explanations) and provide real-time feedback to help recall and synthesize insights.

6.2 Supporting Foraging and Sensemaking

We approached users seeking and deriving explanations through interactive model exploration as engaging in a cognitive process of foraging and sensemaking [27, 83, 162]. Our focus was on helping users update their mental models about how sequential ML models function and what decisions they make. Our findings showed the value of interactions that provide users with a path to find meaningful explanations that match their mental models, expertise, and specific queries. Thus, toolboxes like VIME need to simplify the evidence-gathering process to help users develop and amplify their information-foraging competencies and skills. This will, in turn, allow them to interactively seek evidence and justify relationships between different model inputs and outputs.

Future interactions could allow users to trace decision pathways in the evidence foraging stage by highlighting parts of the Sankey visualization and gathering user feedback to ensure explanations align with model decisions. This could lead to alternative explanation tools, such as summary-based explanations with natural language processing (NLP) for translating model-specific user queries into "what-if" questions. Additionally, caution messages could communicate the model's knowledge limits, clarifying that explanations are reliable only under specific conditions and with adequate output confidence. This could further support users in assigning meaning to the evidence collected through foraging.

6.3 Designing for Broader Stakeholders

Our findings showed that VIME can help computer science-savvy model engineers in their evaluations. Yet, their lack of domain knowledge hindered their ability to suggest solutions for domain-specific model limitations they identified. Such lack of domain and task expertise [120, 151, 169, 174] could further impact the ability of the user to identify strengths and limitations (i.e., trustworthiness) of the models they are exploring [16].

Also, we can not expect model engineers always to be present to debug and monitor their models after deployment. Other stakeholders with diverse domain and task expertise [47, 76, 95, 164, 170] seek to independently explore the models they interact with. Those end-users include domain experts knowledgeable about the model's application area, policymakers regulating those models, and consumers who interact with interfaces supported by those ML models.

Therefore, it is important to evaluate the effectiveness of toolboxes like VIME to support a broader set of stakeholders in ML model evaluation. Formative studies with those end-users can help identify their needs, challenges, and context of use during ML evaluation. Such insights could inform the design of future explanation tools tailored to their needs and expertise.

7 CONCLUSION AND FUTURE WORK

We presented Visual Interactive Model Explorer (VIME), a data-, domain-, and model-agnostic toolbox allowing model engineers to holistically investigate their trained sequential ML models. As a system design contribution [184] towards human-centered eXplainable AI (HCXAI) [48], VIME facilitates interactive exploration, debugging, and monitoring of ML model errors and limitations. Visualizations and interactions in VIME help users effectively explore sequential data and models by decomposing long sequences of model decisions into simple "what-if" scenarios for comparison.

Our findings showed that VIME makes it easier for model engineers to investigate *how* (e.g., how their ML model makes a particular decision?) and *what-if* (e.g., if the input was different, would their ML model make a different decision?). A series of interactions in VIME helped users identify and debug sequential ML model limitations, including over-generalization, spurious correlations, missing feature values, and missing input-output relationships.

Future work should explore the design of explanation tools and toolboxes that target other end-users who may not have computer science-savvy expertise but seek explanations about ML models they interact with. Insights from our iterative design and evaluation of VIME will inform the design of interactive XAI tools that assist diverse stakeholders in evaluating sequential decision-support systems for their broader adoption in real-world domains.

ACKNOWLEDGMENTS

This article is based upon work in part supported by the National Science Foundation under Grant No. IIS-2237562 and Department of Navy award N00014-23-1-2735 issued by the Office of Naval Research. Toyota Research Institute (TRI) provided funds to assist the authors with their research, but this article solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity.

We thank Anna Kratz for providing the MS dataset and selecting relevant "what-if" scenarios for this domain that we used in our evaluation study. We also thank Richard Lewis and Eytan Adar for helping us ground our ideas in the context of existing cognitive science and interactive visualization work. We thank the members of the Computational HCI (CompHCI) Lab at the University of Michigan for brainstorming ideas and proofreading the paper.

REFERENCES

[1] Kjersti Aas, Martin Jullum, and Anders Løland. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence* 298 (2021), 103502. https://doi.org/10.1016/j.artint.2021.103502

[2] Abdallah Abbas, Ciara O'Byrne, Dun Jack Fu, Gabriella Moraes, Konstantinos Balaskas, Robbert Struyven, Sara Beqiri, Siegfried K. Wagner, Edward Korot, and Pearse A. Keane. 2022. Evaluating an automated machine learning model that predicts visual acuity outcomes in patients with neovascular age-related macular degeneration. *Graefe's Archive for Clinical and Experimental Ophthalmology* 260, 8 (01 Aug 2022), 2461–2473. https://doi.org/10.1007/s00417-021-05544-y

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[4] Abdulaziz Alaboudi and Thomas D. Latoza. 2023. Hypothesizer: A Hypothesis-Based Debugger to Find and Test Debugging Hypotheses. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 73, 14 pages. https://doi.org/10.1145/3586183.3606781

[5] Daniel Alcaide and Jan Aerts. 2018. Multilevel Visual Clustering Exploration for Incomplete Time-Series in Water Samples. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 116–117. https://doi.org/10.1109/VAST.2018.8802480

[6] Gulsum Alicioglu and Bo Sun. 2022. A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics* 102 (2022), 502–520. https://doi.org/10.1016/j.cag.2021.09.002

[7] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[8] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 337–346. https://doi.org/10.1145/2702123.2702509

[9] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: Interactive Machine Learning for on-Demand Group Creation in Social Networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2207676.2207680

[10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. In *Ethics of Data and Analytics*. 254–264. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[11] Anindya Das Antar, Anna Kratz, and Nikola Banovic. 2023. Behavior Modeling Approach for Forecasting Physical Functioning of People with Multiple Sclerosis. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 7 (mar 2023), 29 pages. https://doi.org/10.1145/3580887

[12] Sriram Karthik Badam, Jieqiong Zhao, Shivalik Sen, Niklas Elmqvist, and David Ebert. 2016. TimeFork: Interactive Prediction of Time Series. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5409–5420. https://doi.org/10.1145/2858036.2858150

[13] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *The Journal of Machine Learning Research* 11 (2010), 1803–1831.

[14] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey. 2016. Modeling and Understanding Human Routine Behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 248–260. https://doi.org/10.1145/2858036.2858557

[15] Nikola Banovic, Anqi Wang, Yanfeng Jin, Christie Chang, Julian Ramos, Anind K. Dey, and Jennifer Mankoff. 2017. Leveraging Human Routine Models to Detect and Generate Human Behaviors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver, CO, USA, May 06-11, 2017*, Gloria Mark, Susan R. Fussell, Cliff Lampe, m. c. schraefel, Juan Pablo Hourcade, Caroline Appert, and Daniel Wigdor (Eds.). ACM, 6683–6694. https://doi.org/10.1145/3025453.3025571

[16] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (apr 2023), 17 pages. https://doi.org/10.1145/3579460

[17] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[18] Leilani Battle. 2022. Behavior-Driven Testing of Big Data Exploration Tools. *Interactions* 29, 5 (aug 2022), 9–10. https://doi.org/10.1145/3554726

[19] Leilani Battle and Jeffrey Heer. 2019. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau. In *Computer graphics forum*, Vol. 38. Wiley Online Library, 145–159.

[20] Alex Bäuerle, Ángel Alexander Cabrera, Fred Hohman, Megan Maher, David Koski, Xavier Suau, Titus Barik, and Dominik Moritz. 2022. Symphony: Composing Interactive Interfaces for Machine Learning. In *CHI Conference on Human Factors in Computing Systems*. 1–14.

[21] João Bento, Pedro Saleiro, André F. Cruz, Mário A.T. Figueiredo, and Pedro Bizarro. 2021. TimeSHAP: Explaining Recurrent Models through Sequence Perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery &amp; Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2565–2573. https://doi.org/10.1145/3447548.3467166

[22] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On Selective, Mutable and Dialogic XAI: A Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 411, 21 pages. https://doi.org/10.1145/3544548.3581314

[23] Greenwell Brandon M., Boehmke Bradley C., and McCarthy Andrew J. 2018. A Simple and Effective Model-Based Variable Importance Measure. *ArXiv* abs/1805.04755 (2018).

[24] Cynthia A. Brewer. 1994. *Color Use Guidelines for Mapping and Visualization* (c ed.). Number C in Modern Cartography Series. 123–147. https://doi.org/10.

1016/B978-0-08-042415-6.50014-4

[25] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-Making in Child Welfare Services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300271

[26] Ángel Alexander Cabrera, Erica Fu, Donald Bertucci, Kenneth Holstein, Ameet Talwalkar, Jason I. Hong, and Adam Perer. 2023. Zeno: An Interactive Framework for Behavioral Evaluation of Machine Learning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 419, 14 pages. https://doi.org/10.1145/3544548.3581268

[27] Ángel Alexander Cabrera, Marco Tulio Ribeiro, Bongshin Lee, Rob DeLine, Adam Perer, and Steven M. Drucker. 2022. What Did My AI Learn? How Data Scientists Make Sense of Model Behavior. *ACM Trans. Comput.-Hum. Interact.* (may 2022). https://doi.org/10.1145/3542921 Just Accepted.

[28] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[29] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 1721–1730. https://doi.org/10.1145/2783258.2788613

[30] Gromit Yeuk-Yin Chan, Jun Yuan, Kyle Overton, Brian Barr, Kim Rees, Luis Gustavo Nonato, Enrico Bertini, and Cláudio T. Silva. 2020. SUBPLEX: Towards a Better Understanding of Black Box Model Explanations at the Subpopulation Level. *ArXiv* abs/2007.10609 (2020). https://api.semanticscholar.org/CorpusID:220665477

[31] Cheng Chen and S. Shyam Sundar. 2023. Is This AI Trained on Credible Data? The Effects of Labeling Quality and Performance Bias on User Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 816, 11 pages. https://doi.org/10.1145/3544548.3580805

[32] Hongruyu Chen, Fernando Gonzalez, Oto Mraz, Sophia Kuhn, Cristina Guzman, and Mennatallah El-Assady. 2023. Explore, Compare, and Predict Investment Opportunities through What-If Analysis: US Housing Market Investigation. In *Proceedings of the 16th International Symposium on Visual Information Communication and Interaction* (Guangzhou, China) *(VINCI '23)*. Association for Computing Machinery, New York, NY, USA, Article 19, 5 pages. https://doi.org/10.1145/3615522.3615542

[33] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *arXiv preprint arXiv:2301.07255* (2023).

[34] Furui Cheng, Yao Ming, and Huamin Qu. 2021. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2021), 1438–1447. https://doi.org/10.1109/TVCG.2020.3030342

[35] Minseok Cho, Gyeongbok Lee, and Seung-won Hwang. 2019. Explanatory and Actionable Debugging for Machine Learning: A TableQA Demonstration. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Paris, France) *(SIGIR'19)*. Association for Computing Machinery, New York, NY, USA, 1333–1336. https://doi.org/10.1145/3331184.3331404

[36] Michael Chromik and Andreas Butz. 2021. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In *Human-Computer Interaction – INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30 – September 3, 2021, Proceedings, Part II* (Bari, Italy). Springer-Verlag, Berlin, Heidelberg, 619–640. https://doi.org/10.1007/978-3-030-85616-8_36

[37] Michael Correll and Michael Gleicher. 2014. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2142–2151. https://doi.org/10.1109/TVCG.2014.2346298

[38] Tommy Dang, Huyen N. Nguyen, and Ngan V.T. Nguyen. 2021. VixLSTM: Visual Explainable LSTM for Multivariate Time Series. In *The 12th International Conference on Advances in Information Technology* (Bangkok, Thailand) *(IAIT2021)*. Association for Computing Machinery, New York, NY, USA, Article 34, 5 pages. https://doi.org/10.1145/3468784.3471603

[39] Tommy Dang, Hao Van, Huyen Nguyen, Vung Pham, and Rattikorn Hewett. 2020. DeepVix: Explaining Long Short-Term Memory Network With High Dimensional Time Series Data. In *Proceedings of the 11th International Conference on Advances in Information Technology* (Bangkok, Thailand) *(IAIT '20)*. Association for Computing Machinery, New York, NY, USA, Article 38, 10 pages. https://doi.org/10.1145/3406601.3406643

[40] Sanjeeb Dash, Oktay Günlük, and Dennis Wei. 2018. Boolean Decision Rules via Column Generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 4660–4670.

[41] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376638

[42] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. https://doi.org/10.1145/3617694.3623261

[43] Anind K. Dey, Raffay Hamid, Chris Beckmann, Ian Li, and Daniel Hsu. 2004. A CAPpella: Programming by Demonstration of Context-Aware Applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 33–40. https://doi.org/10.1145/985692.985697

[44] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[45] Ruofei Du, Na Li, Jing Jin, Michelle Carney, Xiuxiu Yuan, Kristen Wright, Mark Sherwood, Jason Mayes, Lin Chen, Jun Jiang, Jingtao Zhou, Zhongyi Zhou, Ping Yu, Adarsh Kowdle, Ram Iyengar, and Alex Olwal. 2023. Experiencing Visual Blocks for ML: Visual Prototyping of AI Pipelines. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 76, 3 pages. https://doi.org/10.1145/3586182.3615817

[46] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, and Rajiv Ranjan. 2023. Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Comput. Surv.* 55, 9, Article 194 (jan 2023), 33 pages. https://doi.org/10.1145/3561048

[47] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. https://doi.org/10.48550/ARXIV.2107.13509

[48] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 109, 7 pages. https://doi.org/10.1145/3491101.3503727

[49] Jerry Fails and Dan Olsen. 2003. A Design Tool for Camera-Based Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Ft. Lauderdale, Florida, USA) *(CHI '03)*. Association for Computing Machinery, New York, NY, USA, 449–456. https://doi.org/10.1145/642611.642690

[50] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (Miami, Florida, USA) *(IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. https://doi.org/10.1145/604045.604056

[51] Juliana J. Ferreira and Mateus S. Monteiro. 2020. What Are People Doing About XAI User Experience? A Survey on AI Explainability Research and Practice. In *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, Aaron Marcus and Elizabeth Rosenzweig (Eds.). Springer International Publishing, Cham, 56–73.

[52] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: Interactive Concept Learning in Image Search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/1357054.1357061

[53] Jules Françoise, Baptiste Caramiaux, and Téo Sanchez. 2021. Marcelle: Composing Interactive Machine Learning Workflows and Interfaces. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 39–53. https://doi.org/10.1145/3472749.3474734

[54] Jerome H Friedman and Bogdan E Popescu. 2008. Predictive learning via rule ensembles. *The annals of applied statistics* (2008), 916–954.

[55] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R. Millen, Murray Campbell, Sadhana Kumaravel, and Wei Zhang. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376316

[56] Shilpa Gite, Hrituja Khatavkar, Ketan Kotecha, Shilpi Srivastava, Priyam Maheshwari, and Neerav Pandey. 2021. Explainable stock prices prediction from financial news articles using sentiment analysis. *PeerJ Comput. Sci.* 7, e340 (Jan.

2021), e340.

[57] Stefan Grafberger, Paul Groth, and Sebastian Schelter. 2022. Towards Data-Centric What-If Analysis for Native Machine Learning Pipelines. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning* (Philadelphia, Pennsylvania) *(DEEM '22)*. Association for Computing Machinery, New York, NY, USA, Article 3, 5 pages. https://doi.org/10.1145/3533028.3533303

[58] Stefan Grafberger, Paul Groth, and Sebastian Schelter. 2023. Automating and Optimizing Data-Centric What-If Analyses on Native Machine Learning Pipelines. *Proc. ACM Manag. Data* 1, 2, Article 128 (jun 2023), 26 pages. https://doi.org/10.1145/3589273

[59] Ben Green. 2022. The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review* 45 (2022), 105681. https://doi.org/10.1016/j.clsr.2022.105681

[60] Tovi Grossman and George Fitzmaurice. 2015. An Investigation of Metrics for the In Situ Detection of Software Expertise. *Human–Computer Interaction* 30, 1 (2015), 64–102. https://doi.org/10.1080/07370024.2014.881668

[61] Damodar N. Gujarati and Dawn C. Porter. 2009. *Basic Econometrics. Chapter: The Nature and Sources of Data for Economic Analysis* (fifth international ed.). McGraw-Hill, New York. 22–28 pages.

[62] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science Robotics* 4, 37 (2019), eaay7120. https://doi.org/10.1126/scirobotics.aay7120

[63] Yi Guo, Shunan Guo, Zhuochen Jin, Smiti Kaul, David Gotz, and Nan Cao. 2022. Survey on visual analysis of event sequence data. *IEEE Trans. Vis. Comput. Graph.* 28, 12 (Dec. 2022), 5091–5112. https://doi.org/10.1109/TVCG.2021.3100413

[64] Sophia Hadash, Martijn C. Willemsen, Chris Snijders, and Wijnand A. IJsselsteijn. 2022. Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 487, 9 pages. https://doi.org/10.1145/3491102.3517650

[65] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (nov 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[66] Degan Hao, Lei Zhang, Jules Sumkin, Aly Mohamed, and Shandong Wu. 2020. Inaccurate labels in weakly-supervised deep learning: Automatic identification and correction and their impact on classification performance. *IEEE J. Biomed. Health Inform.* 24, 9 (Sept. 2020), 2701–2710.

[67] Mark Harrower and Cynthia A. Brewer. 2003. ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps. *The Cartographic Journal* 40, 1 (2003), 27–37. https://doi.org/10.1179/000870403235002042

[68] Björn Hartmann, Leith Abdulla, Manas Mittal, and Scott R. Klemmer. 2007. Authoring Sensor-Based Interactions by Demonstration with Direct Manipulation and Pattern Recognition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/1240624.1240646

[69] Frederick Hayes-Roth. 1985. Rule-Based Systems. *Commun. ACM* 28, 9 (sep 1985), 921–932. https://doi.org/10.1145/4284.4286

[70] Harry Hochheiser and Ben Shneiderman. 2001. Interactive Exploration of Time Series Data. In *Proceedings of the 4th International Conference on Discovery Science (DS '01)*. Springer-Verlag, Berlin, Heidelberg, 441–446.

[71] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[72] Jin-Hyuk Hong, Ben Margines, and Anind K. Dey. 2014. A Smartphone-Based Sensing Platform to Model Aggressive Driving Behaviors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4047–4056. https://doi.org/10.1145/2556288.2557321

[73] Mir Riyanul Islam, Mobyen Uddin Ahmed, Shaibal Barua, and Shahina Begum. 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences* 12, 3 (2022), 1353.

[74] Abhinav Jain, Hima Patel, Lokesh Nagalapatti, Nitin Gupta, Sameep Mehta, Shanmukha Guttula, Shashank Mujumdar, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2020. Overview and Importance of Data Quality for Machine Learning Tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 3561–3562. https://doi.org/10.1145/3394486.3406477

[75] Tony Jebara and Marina Meila. 2006. Machine learning: Discriminative and generative. *The Mathematical Intelligencer* 28 (03 2006), 67–69. https://doi.org/10.1007/BF02987011

[76] Jinglu Jiang, Surinder Kahai, and Ming Yang. 2022. Who Needs Explanation and When? Juggling Explainable AI and User Epistemic Uncertainty. *Int. J. Hum.-Comput. Stud.* 165, C (sep 2022), 17 pages. https://doi.org/10.1016/j.ijhcs.2022.102839

[77] Sara Johansson. 2009. Visual Exploration of Categorical and Mixed Data Sets. In *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery: Integrating Automated Analysis with Interactive Exploration* (Paris, France) *(VAKD '09)*. Association for Computing Machinery, New York, NY, USA, 21–29. https://doi.org/10.1145/1562849.1562852

[78] Anker Helms Jorgensen. 1990. Thinking-aloud in user interface design: A method promoting cognitive ergonomics. *Ergonomics* 33 (04 1990). https://doi.org/10.1080/00140139008927157

[79] Ericsson K. Anders and Simon Herbert A. 1984. Protocol Analysis: Verbal Reports as Data.

[80] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Chau. 2018. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics* 24 (2018), 88–97.

[81] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 895–905.

[82] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2022. A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations. *ACM Comput. Surv.* 55, 5, Article 95 (dec 2022), 29 pages. https://doi.org/10.1145/3527848

[83] Harmanpreet Kaur, Eytan Adar, Eric Gilbert, and Cliff Lampe. 2022. Sensible AI: Re-imagining Interpretability and Explainability using Sensemaking Theory. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 702–714. https://doi.org/10.1145/3531146.3533135

[84] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. *Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376219

[85] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3819–3828. https://doi.org/10.1145/2702123.2702520

[86] Antino Kim, Mochen Yang, and Jingjing Zhang. 2023. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *ACM Trans. Comput.-Hum. Interact.* 30, 1, Article 14 (mar 2023), 36 pages. https://doi.org/10.1145/3557889

[87] Andy Kirk. 2016. *Data Visualisation: A Handbook for Data Driven Design*. Sage Publications Ltd.

[88] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[89] Alina Köchling and Marius Claus Wehner. 2020. Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13, 3 (01 Nov 2020), 795–848. https://doi.org/10.1007/s40685-020-00134-w

[90] Anna L Kratz, Tiffany J Braley, Emily Foxen-Craft, Eric Scott, John F Murphy III, and Susan L Murphy. 2017. How do pain, fatigue, depressive, and cognitive symptoms relate to well-being and social and physical functioning in the daily lives of individuals with multiple sclerosis? *Archives of physical medicine and rehabilitation* 98, 11 (2017), 2160–2166.

[91] Andreas Krause and Carlos Guestrin. 2005. Near-Optimal Nonmyopic Value of Information in Graphical Models. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence* (Edinburgh, Scotland) *(UAI'05)*. AUAI Press, Arlington, Virginia, USA, 324–331.

[92] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5686–5697. https://doi.org/10.1145/2858036.2858529

[93] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[94] Michelle S. Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A. Landay, and Michael S. Bernstein. 2023. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 741, 24 pages. https://doi.org/10.1145/3544548.3581290

[95] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*

296 (2021), 103473. https://doi.org/10.1016/j.artint.2021.103473

[96] Chi-Hoon Lee. 2010. Learning to Combine Discriminative Classifiers: Confidence Based. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Washington, DC, USA) *(KDD '10)*. Association for Computing Machinery, New York, NY, USA, 743–752. https://doi.org/10.1145/1835804.1835899

[97] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[98] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 699, 13 pages. https://doi.org/10.1145/3411764.3445261

[99] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[100] Le Liu, Lace Padilla, Sarah H. Creem-Regehr, and Donald H. House. 2019. Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 882–891. https://doi.org/10.1109/TVCG.2018.2865193

[101] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[102] Scott M Lundberg and Su-In Lee. 2018. Consistent individualized feature attribution for tree ensembles. In *Advances in neural information processing systems*. 6469–6480.

[103] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 759, 19 pages. https://doi.org/10.1145/3544548.3581058

[104] Michael Madaio, Shang-Tse Chen, Oliver L. Haimson, Wenwen Zhang, Xiang Cheng, Matthew Hinds-Aldrich, Duen Horng Chau, and Bistra Dilkina. 2016. Firebird: Predicting Fire Risk and Prioritizing Fire Inspections in Atlanta. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 185–194. https://doi.org/10.1145/2939672.2939682

[105] Sana Malik, Fan Du, Megan Monroe, Eberechukwu Onukwugha, Catherine Plaisant, and Ben Shneiderman. 2015. Cohort Comparison of Event Sequences with Balanced Integration of Visual Analytics and Statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (Atlanta, Georgia, USA) *(IUI '15)*. Association for Computing Machinery, New York, NY, USA, 38–49. https://doi.org/10.1145/2678025.2701407

[106] Sana Malik, Ben Shneiderman, Fan Du, Catherine Plaisant, and Margret Bjarnadottir. 2016. High-Volume Hypothesis Testing: Systematic Exploration of Event Sequence Comparisons. *ACM Trans. Interact. Intell. Syst.* 6, 1, Article 9 (mar 2016), 23 pages. https://doi.org/10.1145/2890478

[107] Keri Mallari, Kori Inkpen, Paul Johns, Sarah Tan, Divya Ramesh, and Ece Kamar. 2020. *Do I Look Like a Criminal? Examining How Race Presentation Impacts Human Judgement of Recidivism.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376257

[108] Masayoshi Mase, Art B Owen, and Benjamin Seiler. 2019. Explaining black box decisions by shapley cohort refinement. *arXiv preprint arXiv:1911.00467* (2019).

[109] Afra Mashhadi, Annuska Zolyomi, and Jay Quedado. 2022. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 37, 7 pages. https://doi.org/10.1145/3491101.3503568

[110] Dan Maynes-Aminzade, Terry Winograd, and Takeo Igarashi. 2007. Eyepatch: Prototyping Camera-Based Interaction through Examples. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology* (Newport, Rhode Island, USA) *(UIST '07)*. Association for Computing Machinery, New York, NY, USA, 33–42. https://doi.org/10.1145/1294211.1294219

[111] Sean McGregor, Hailey Buckingham, Rachel Houtman, Claire A Montgomery, Ronald A Metoyer, and Thomas G Dietterich. 2015. MDPVIS: An Interactive Visualization for Testing Markov Decision Processes.. In *AAAI Fall Symposia*. 56–58.

[112] Roisin McNaney, Catherine Morgan, Pranav Kulkarni, Julio Vega, Farnoosh Heidarivincheh, Ryan McConville, Alan Whone, Mickey Kim, Reuben Kirkham, and Ian Craddock. 2022. Exploring Perceptions of Cross-Sectoral Data Sharing with People with Parkinson's. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 280, 14 pages. https://doi.org/10.1145/3491102.3501984

[113] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, 2 (April 2008), 194–210. https://doi.org/10.1518/001872008x288574

[114] Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. 2020. Importance Sampling Techniques for Policy Optimization. *J. Mach. Learn. Res.* 21, 1, Article 141 (jan 2020), 75 pages.

[115] Yao Ming, Panpan Xu, Furui Cheng, Huamin Qu, and Liu Ren. 2020. ProtoSteer: Steering Deep Sequence Model with Prototypes. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 238–248. https://doi.org/10.1109/TVCG.2019.2934267

[116] T.P. Minka and R.W. Picard. 1996. Interactive learning with a "Society of Models". In *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 447–452. https://doi.org/10.1109/CVPR.1996.517110

[117] Michael Mitzenmacher and Sergei Vassilvitskii. 2022. Algorithms with Predictions. *Commun. ACM* 65, 7 (jun 2022), 33–35. https://doi.org/10.1145/3528087

[118] Katelyn Morrison, Mayank Jain, Jessica Hammer, and Adam Perer. 2023. Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 273 (oct 2023), 22 pages. https://doi.org/10.1145/3610064

[119] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 48 (apr 2023), 37 pages. https://doi.org/10.1145/3579481

[120] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 183 (apr 2024), 39 pages. https://doi.org/10.1145/3641022

[121] Eduardo Mosqueira-Rey, Elena Hernández Pereira, David Alonso-Ríos, and José Bobes-Bascarán. 2022. A Classification and Review of Tools for Developing and Interacting with Machine Learning Systems. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing* (Virtual Event) *(SAC '22)*. Association for Computing Machinery, New York, NY, USA, 1092–1101. https://doi.org/10.1145/3477314.3507310

[122] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.

[123] Maath Musleh, Angelos Chatzimparmpas, and Ilir Jusufi. 2021. Visual Analysis of Industrial Multivariate Time Series. In *Proceedings of the 14th International Symposium on Visual Information Communication and Interaction* (Potsdam, Germany) *(VINCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 3, 5 pages. https://doi.org/10.1145/3481549.3481557

[124] Yuri Nakao, Simone Stumpf, Subeida Ahmed, Aisha Naseer, and Lorenzo Strappelli. 2022. Toward Involving End-users in Interactive Human-in-the-loop AI Fairness. *ACM Trans. Interact. Intell. Syst.* 12, 3, Article 18 (jul 2022), 30 pages. https://doi.org/10.1145/3514258

[125] George E. Newman and Brian J. Scholl. 2012. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review* 19, 4 (01 Aug 2012), 601–607. https://doi.org/10.3758/s13423-012-0247-5

[126] Jakob Nielsen. 1989. Usability Engineering at a Discount. In *Proceedings of the Third International Conference on Human-Computer Interaction on Designing and Using Human-Computer Interfaces and Knowledge Based Systems (2nd Ed.)* (Boston, Massachusetts, USA). Elsevier Science Inc., USA, 394–401.

[127] Jakob Nielsen. 2012. *How Many Test Users in a Usability Study?* https://www.nngroup.com/articles/how-many-test-users/ Accessed: 07/01/2023.

[128] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. https://doi.org/10.1145/3397481.3450639

[129] Heru Nugroho and Kridanto Surendro. 2019. Missing Data Problem in Predictive Analytics. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications* (Penang, Malaysia) *(ICSCA '19)*. Association for Computing Machinery, New York, NY, USA, 95–100. https://doi.org/10.1145/3316615.3316730

[130] Google PAIR. 2022. *Facet Dive?* https://pair-code.github.io/facets/.

[131] Reshika Palaniyappan Velumani, Meng Xia, Jun Han, Chaoli Wang, ALEXIS K LAU, and Huamin Qu. 2022. AQX: Explaining Air Quality Forecast for Verifying Domain Knowledge Using Feature Importance Visualization. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 720–733.

https://doi.org/10.1145/3490099.3511150

[132] Linsey Pang, Wei Liu, Lingfei Wu, Kexin Xie, Stephen Guo, Raghav Chalapathy, and Musen Wen. 2022. Applied Machine Learning Methods for Time Series Forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (Atlanta, GA, USA) *(CIKM '22)*. Association for Computing Machinery, New York, NY, USA, 5175–5176. https://doi.org/10.1145/3511808.3557492

[133] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How Accurate Does It Feel? – Human Perception of Different Types of Classification Mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 180, 13 pages. https://doi.org/10.1145/3491102.3501915

[134] Chunjong Park, Anas Awadalla, Tadayoshi Kohno, and Shwetak Patel. 2021. Reliable and Trustworthy Machine Learning for Health Using Dataset Shift Detection. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 3043–3056. https://proceedings.neurips.cc/paper/2021/file/17e23e50bedc63b4095e3d8204ce063b-Paper.pdf

[135] Kayur Patel, Naomi Bancroft, Steven M. Drucker, James Fogarty, Amy J. Ko, and James Landay. 2010. Gestalt: Integrated Support for Implementation and Analysis in Machine Learning. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. Association for Computing Machinery, New York, NY, USA, 37–46. https://doi.org/10.1145/1866029.1866038

[136] Adam Perer and David Gotz. 2013. Data-Driven Exploration of Care Plans for Patients. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 439–444. https://doi.org/10.1145/2468356.2468434

[137] Adam Perer and Fei Wang. 2014. Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences. In *Proceedings of the 19th International Conference on Intelligent User Interfaces* (Haifa, Israel) *(IUI '14)*. Association for Computing Machinery, New York, NY, USA, 153–162. https://doi.org/10.1145/2557500.2557508

[138] Daniel Petrov, Rakan Alseghayer, Mohamed Sharaf, Panos K. Chrysanthis, and Alexandros Labrinidis. 2017. Interactive Exploration of Correlated Time Series. In *Proceedings of the ExploreDB'17* (Chicago, IL, USA) *(ExploreDB'17)*. Association for Computing Machinery, New York, NY, USA, Article 2, 6 pages. https://doi.org/10.1145/3077331.3077335

[139] P. Jonathon Phillips, Carina Hahn, Peter Fontana, Amy Yates, Kristen Greene, David Broniatowski, and Mark Przybocki. 2021. Four Principles of Explainable Artificial Intelligence. NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD. https://doi.org/10.6028/NIST.IR.8312

[140] David Piorkowski, Inge Vejsbjerg, Owen Cornec, Elizabeth M. Daly, and Öznur Alkan. 2023. AIMEE: An Exploratory Study of How Rules Support AI Developers to Explain and Edit Models. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 255 (oct 2023), 25 pages. https://doi.org/10.1145/3610046

[141] Snehal Prabhudesai, Nicholas Chandler Wang, Vinayak Ahluwalia, Xun Huan, Jayapalli Rajiv Bapuraj, Nikola Banovic, and Arvind Rao. 2021. Stratification by Tumor Grade Groups in a Holistic Evaluation of Machine Learning for Brain Tumor Segmentation. *Frontiers in Neuroscience* 15 (2021), 1236. https://doi.org/10.3389/fnins.2021.740353

[142] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. https://doi.org/10.1145/3581641.3584033

[143] Maria Priestley, Fionntán O'donnell, and Elena Simperl. 2023. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *J. Data and Information Quality* 15, 2, Article 11 (jun 2023), 39 pages. https://doi.org/10.1145/3592616

[144] Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[145] JR Quinlan. 1986. Induction of decision trees. mach. learn. (1986).

[146] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[147] P. Riehmann, M. Hanfler, and B. Froehlich. 2005. Interactive Sankey diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* 233–240. https://doi.org/10.1109/INFVIS.2005.1532152

[148] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. 2021. Explainable Artificial Intelligence (XAI) on TimeSeries Data: A Survey. arXiv:2104.00950 [cs.LG]

[149] Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L. Glassman, and Finale Doshi-Velez. 2021. Evaluating the Interpretability of Generative Models by Interactive Reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 80, 15 pages. https://doi.org/10.1145/3411764.3445296

[150] Martin Rosvall and Carl T Bergstrom. 2010. Mapping change in large networks. *PLoS One* 5, 1 (2010). https://doi.org/10.1371/journal.pone.0008694

[151] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A Missing Piece in the Puzzle: Considering the Role of Task Complexity in Human-AI Decision Making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) *(UMAP '23)*. Association for Computing Machinery, New York, NY, USA, 215–227. https://doi.org/10.1145/3565472.3592959

[152] Udo Schlegel, Daniela Oelke, Daniel A. Keim, and Mennatallah El-Assady. 2023. Visual Explanations with Attributions and Counterfactuals on Time Series Classification. arXiv:2307.08494 [cs.HC] https://arxiv.org/pdf/2307.08494.pdf

[153] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1330–1340. https://doi.org/10.1145/3531146.3533189

[154] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. https://doi.org/10.1109/VL.1996.545307

[155] Eleanor M Simonsick, Jennifer A Schrack, Nancy W Glynn, and Luigi Ferrucci. 2014. Assessing fatigability in mobility-intact older adults. *Journal of the American Geriatrics Society* 62, 2 (feb 2014), 347–351.

[156] Torty Sivill and Peter Flach. 2022. LIMESegment: Meaningful, Realistic Time Series Explanations. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, 3418–3433. https://proceedings.mlr.press/v151/sivill22a.html

[157] Terry A Slocum, Robert B McMaster, Fritz C Kessler, and Hugh H Howard. 2022. *Thematic cartography and geovisualization*. CRC Press.

[158] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[159] Angelo Sotgiu, Maura Pintor, and Battista Biggio. 2022. Explainability-Based Debugging of Machine Learning for Vulnerability Discovery. In *Proceedings of the 17th International Conference on Availability, Reliability and Security* (Vienna, Austria) *(ARES '22)*. Association for Computing Machinery, New York, NY, USA, Article 113, 8 pages. https://doi.org/10.1145/3538969.3543809

[160] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M. Rush. 2019. Seq2seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan 2019), 353–363. https://doi.org/10.1109/TVCG.2018.2865044

[161] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. 2018. LSTMVis: A Tool for Visual Analysis of Hidden State Dynamics in Recurrent Neural Networks. *IEEE Transactions on Visualization &amp; Computer Graphics* 24, 01 (jan 2018), 667–676. https://doi.org/10.1109/TVCG.2017.2744158

[162] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. https://doi.org/10.1145/3586183.3606756

[163] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-Based Design. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 212–228. https://doi.org/10.1145/3490099.3511119

[164] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and Their Needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 74, 16 pages. https://doi.org/10.1145/3411764.3445088

[165] Martin Theus. 2002. *Data Visualization for Domain Exploration: Highly Multivariate Interaction Techniques*. Oxford University Press, Inc., USA, 232–241.

[166] Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. 2023. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence* 5, 3 (01 Mar 2023), 250–260. https://doi.org/10.1038/

s42256-023-00620-w

[167] Niels van Berkel and Kasper Hornbæk. 2023. Implications of Human-Computer Interaction Research. *Interactions* 30, 4 (jun 2023), 50–55. https://doi.org/10.1145/3600103

[168] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 251, 17 pages. https://doi.org/10.1145/3544548.3580903

[169] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[170] Beatrice Vincenzi, Simone Stumpf, Alex S. Taylor, and Yuri Nakao. 2024. Lay User Involvement in Developing Human-Centric Responsible AI Systems: When and How? *ACM J. Responsib. Comput.* (mar 2024). https://doi.org/10.1145/3652592 Just Accepted.

[171] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[172] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[173] Xumeng Wang, Wei Chen, Jiazhi Xia, Zexian Chen, Dongshi Xu, Xiangyang Wu, Mingliang Xu, and Tobias Schreck. 2020. ConceptExplorer: Visual Analysis of Concept Drifts in Multi-source Time-series Data. In *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 1–11. https://doi.org/10.1109/VAST50239.2020.00006

[174] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[175] Xinru Wang and Ming Yin. 2022. Effects of Explanations in AI-Assisted Decision Making: Principles and Comparisons. *ACM Trans. Interact. Intell. Syst.* 12, 4, Article 27 (nov 2022), 36 pages. https://doi.org/10.1145/3519266

[176] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 758, 19 pages. https://doi.org/10.1145/3544548.3581366

[177] Zhenlin Wang, Xun Huan, and Krishna Garikipati. 2019. Variational system identification of the partial differential equations governing the physics of pattern-formation: Inference under varying fidelity and noise. *Computer Methods in Applied Mechanics and Engineering* 356, 1 (nov 2019), 44–74. https://doi.org/10.1016/j.cma.2019.07.007

[178] Zitao Wang, Jun Yan, Yizhou Huang, Jason W Smith, Yu Zhou, Kongyi Cao, Qi Zhang, and Xiang Zhang. 2017. Attend and diagnose: Clinical time series analysis using attention models. In *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1101–1106.

[179] Lindsay Wells and Tomasz Bednarz. 2021. Explainable ai and reinforcement learning—a systematic review of current approaches and trends. *Frontiers in artificial intelligence* 4 (2021), 550030.

[180] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[181] James Wexler, Mahima Pushkarna, Sara Robinson, Tolga Bolukbasi, and Andrew Zaldivar. 2020. Probing ML Models for Fairness with the What-If Tool and SHAP: Hands-on Tutorial. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 705. https://doi.org/10.1145/3351095.3375662

[182] David Wicks. 2017. The coding manual for qualitative researchers. *Qualitative research in organizations and management: an international journal* 12, 2 (2017), 169–170.

[183] M. B. Wilk and R. Gnanadesikan. 1968. Probability Plotting Methods for the Analysis of Data. *Biometrika* 55, 1 (1968), 1–17. http://www.jstor.org/stable/2334448

[184] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions* 23, 3 (apr 2016), 38–44. https://doi.org/10.1145/2907069

[185] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. 2019. Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis. *ACM Trans. Comput.-Hum. Interact.* 26, 4, Article 24 (jun 2019), 27 pages. https://doi.org/10.1145/3319616

[186] Junran Yang, Alex Bäuerle, Dominik Moritz, and Çağatay Demiralp. 2023. VegaProf: Profiling Vega Visualizations. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 94, 11 pages. https://doi.org/10.1145/3586183.3606790

[187] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 585–596. https://doi.org/10.1145/3196709.3196730

[188] Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. 2023. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 248, 15 pages. https://doi.org/10.1145/3544548.3580945

[189] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2021. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 7, 1 (March 2021), 3–36. https://doi.org/10.1007/s41095-020-0191-7

[190] Enhao Zhang and Nikola Banovic. 2021. Method for Exploring Generative Adversarial Networks (GANs) via Automatically Generated Image Galleries. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 76, 15 pages. https://doi.org/10.1145/3411764.3445714

[191] Jinghe Zhang, Kamran Kowsari, James Harrison, Jennifer Lobo, and Laura Barnes. 2018. Patient2Vec: A Personalized Interpretable Deep Representation of the Longitudinal Electronic Health Record. *IEEE Access* (10 2018). https://doi.org/10.1109/ACCESS.2018.2875677

[192] Jiawei Zhang, Yang Wang, Piero Molino, Lezhi Li, and David S. Ebert. 2019. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 25 (2019), 364–373.

[193] Xiong Zhang and Philip J. Guo. 2019. Mallard: Turn the Web into a Contextualized Prototyping Environment for Machine Learning. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 605–618. https://doi.org/10.1145/3332165.3347936

[194] Zhi-Hua Zhou, Yuan Jiang, and Shi-Fu Chen. 2003. Extracting Symbolic Rules from Trained Neural Network Ensembles. *AI Commun.* 16, 1 (jan 2003), 3–15.

[195] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Explaining Machine Learning Models for High-Stakes Decision Making. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 315, 6 pages. https://doi.org/10.1145/3411763.3451743

[196] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2022. Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2022), 1161–1171. https://doi.org/10.1109/TVCG.2021.3114864