Interact: A visual what-if analysis tool for virtual product design

Vasile Ciorna, Guy Melançon, Frank Petry, Mohammad Ghoniem

INTERACT: A Visual What-If Analysis Tool for Virtual Product Design

Information Visualization
2023, Vol.X(X):1–16

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Vasile Ciorna [1,2,3], Guy Melançon [2,3], Frank Petry [1] and Mohammad Ghoniem [4]

Abstract

Virtual prototyping is increasingly used by businesses to streamline operations, cut costs, and enhance daily operations. This often includes a variety of modeling techniques among which, complex, black-box models. The path from model development to utilization in applied contexts is yet long. Domain experts need to be convinced of the validity of the models and to trust their predictions. To be used in the field, model capabilities need to be affordable, i.e., allow rapid and interactive scenario building, even for non-experts. Complex relations governed by statistical interactions must be unveiled for users to understand unexpected predictions. We propose INTERACT, a model-agnostic, visual what-if tool for regression problems, supporting 1) the visualization of statistical interactions between features, 2) the creation of interactive what-if scenarios using predictive models, 3) the evaluation of model quality and building trust, and 4) the externalization of knowledge through model explainability. While the approach applies in various industrial contexts, we validate the application purpose and design with a detailed case study and a qualitative user study with engineers in the tire industry. By unraveling statistical interactions between features, the INTERACT tool proves to be useful to increase the transparency of black-box machine learning models. We also reflect on lessons learned concerning the development of visual what-if tools for virtual product development and beyond.

Keywords

What-if analysis, statistical interactions, design study, visualization

Introduction

Modern organizations seek to exploit the data they collect from their environment and activities, often aiming to consolidate, enrich or improve internal decision-making processes. Modeling techniques can be used to improve these often suboptimal processes, and thus become a companion technique guiding decision-making. Recently, we have seen many examples of the use of Machine Learning (ML) techniques, in various application contexts, to support different processes or tasks. In this article, we focus on the case where models support product design. Designing a product is a decision-making activity, consisting in setting a number of parameters, which will in the end determine the characteristics of the designed product. The decision can sometimes be difficult because desirable characteristics can contradict one another. Flexibility or ease of use, for example, is often the opposite of robustness or durability of a product. This antagonism is reflected at the parameter level: increasing the value of a parameter may affect another one, such as by limiting its range. Such effects may be hard to capture and describe, since parameter interdependencies are widespread across the design space. Human expertise is thus mandatory to assess the physical reality behind these effects, and to help to identify parameter settings that correspond to interesting and feasible solutions. In this article, we focus on the need to exploit many inputs governing a single-output model, therefore deliberately not overlapping with areas such as multi-objective optimization.

Much research has looked at combining state-of-the-art Machine Learning (ML) techniques with visualization tools in various application fields, including healthcare [1,2], meteorology [3,4], and social sciences [5]. Some of this work falls under the umbrella of Explainable Artificial Intelligence (XAI) research [6], insofar that visualization can help to audit and to build trust in the predictions given by AI models. Techniques relying on model-specific and model-agnostic approaches were developed to allow explainability and interpretability [7]. Most efforts in XAI, however, have focused on helping a user to understand a model. Aspects involving practical, but complex, decision-making problems in industrial contexts and, more specifically, in product design are less covered. In this context, application domain experts need to probe models to assess that they faithfully embrace the physical characteristics of the objects being designed. Additionally, they need help to comprehend why some input parameter combinations lead to unexpected outcomes and to recognize the underlying characteristic relationships.

Our approach is inspired by the industrial context of tire design. "Tires are highly engineered structural composites

[1]Goodyear Innovation Center Luxembourg.
[2]University of Bordeaux.
[3]LaBRI: Laboratoire Bordelais de Recherche en Informatique.
[4]Luxembourg Institute of Science and Technology

Corresponding author:
Vasile Ciorna, University of Bordeaux, LaBRI: Laboratoire Bordelais de Recherche en Informatique.
Goodyear Innovation Center Luxembourg.
Email: vasile_ciorna@goodyear.com

whose performance can be designed to meet the vehicle manufacturers' ride, handling, and traction criteria, plus the quality and performance expectations of the customer"[8]. Incidentally, the required characteristics of a tire for a sporty car differ very much from those required from a tire intended for a sedan or electric car. Sporty tires will have a high focus on handling[11] while the others will focus more on acoustics[12] and fuel efficiency[13]. However, from a physics perspective, these disciplines are quite diverse and might be considered as entirely independent sciences[14,15], necessitating various modeling methodologies. This variety, in our opinion, may be encountered in other sectors and is not limited to the design of tires. In a broader context, our approach is relevant whenever a model needs to reflect the physical constraints inherent in a product, whether it is the soles of sports shoes or the casting of steel, for example, requiring experts to be able to assess the model's ability to reflect physical reality.

As an example of an engineering activity in the tire design domain, an acoustics expert needs to understand how a specific tire interacts with the road surface and with the car, and how noise is propagated through air[16] and car structures[17] to the passenger. She needs to understand which design features can be tuned so that the final acoustic comfort inside the car is reached. This task is complex as these features interact, in the statistical sense[18] as we shall explain. Put simply, the influence of a design feature on the outcome depends on one or multiple other features. For instance, a known mechanical relation such as an increase of noise as mass decreases[19] might be invalidated when tire parameters related to stiffness are changed simultaneously. Such relations are at the border of recognized mechanical behaviors and are hard to process, especially in industrial contexts, in time-sensitive conditions. As building and testing physical products is often costly, the engineers need to be supported in virtual assessments.

The final design decision engineers will take often has high financial, production, testing and planning consequences. This is when state-of-the-art methods can help the specialist to de-risk this virtual assessment process.

The contributions of this work are: **1 - INTERACT, a what–if scenario system providing a novel combination of ML-based regression models and statistical interaction analysis.** The system allows the user to: (a) Create and compare what-if scenarios together with a measure of confidence to support forecasting, (b) Understand model predictions by making statistical interactions more affordable; **2 - a case study and a qualitative user study** with participants from the tire industry. **3 - a report of lessons learned concerning desirable characteristics of what-if systems.**

Related Work

INTERACT relates to various areas of research, including visualization and explainable machine learning, statistical interactions, what-if tools and decision making.

Visualization for model understanding

Recently, much work has focused on Explainable Artificial Intelligence (XAI)[20–24]. Chatzimparmpas *et al.*[7] organize prior work on visualization for interpreting ML models in six categories: Visual Analytics Pipelines, General ML models, Predictive Visual Analytics (PVA), Interactive ML (IML), Deep Learning and Dimensionality Reduction. Work in the PVA and IML categories is closer to INTERACT. In another extensive literature review, Chatzimparmpas *et al.*[25] stress the importance of trust in ML and introduce five levels of trust directed to various aspects of IML.

Adadi and Berrada[26] distinguish model-agnostic and model-specifc XAI methods. Model-specific approaches are usually bound to the specific type of model used to fit the data. Examples of model-specific approaches are iForest[27] for Random Forests models, GAN Lab[28] for Generative Adversarial Networks or DQNViztool[29] for Deep Q Networks. Interested readers can refer to additional papers on the visualization of neural networks[1,30–32]. Model-specific solutions can be discounted with the advent of a new, more effective type of ML model.

In INTERACT, we take a model-agnostic approach for regression problems. Model-agnostic approaches are independent of the type of model used to fit the data, which makes them more likely to stand the test of time. As discussed by Spinner *et al.*[33], model-agnostic techniques can be distinguished according to their explanation coverage. An explanation is local when it is valid for a certain data sample, or global when it holds for the entire data set.

LIME[34] uses local surrogate models to explain predictions, e.g., by using words, or parts of an image, that lead to a certain model decision. SHAP[35] provides both local and global explanations. Its force plots depict which features contribute to a single model output, whether positively or negatively, and the global explanations are achieved by rotating these plots and concatenating them horizontally on a complete dataset. A feature overview helps to identify the most important features of the observed model. RuleMatrix[36] uses rule induction to provide explanations to non-ML practitioners who need to understand and use ML models. Going beyond the explanation of one model, the comparative analysis of multiple models has been the focus of Manifold[37] and Clustervision[38]. In INTERACT, we go beyond model-agnostic explainability by supporting the creation of what-if scenarios on the fly, together with an estimation of confidence for each prediction.

Visualization of statistical interactions

Statistical interactions, or interaction effects, are widely studied in statistics[39–42], but rarely in ML-enabled visualization tools. They capture a non-additive influence of two variables to a model outcome. Typically, while variable X may vary monotonously to the dependent variable Y, the rate at which it varies may depend on a second variable Z. The H-statistic is a measure of interaction effects over the data distribution, proportional to the power of the interaction, but it may yield spurious interactions[43]. Partial Dependence (PD) models are also used to evaluate interaction effects[44], without raising spurious interactions. SHAP[35] evaluates interactions using SHAP values. Save for a few exceptions[45], the H-statistic and the PD-based method were not often used to uncover interaction effects in ML models. Interaction effects are visualized using line charts, such as Individual Conditional Expectation (ICE) lines[46] and SHAP plots[35]. Individual

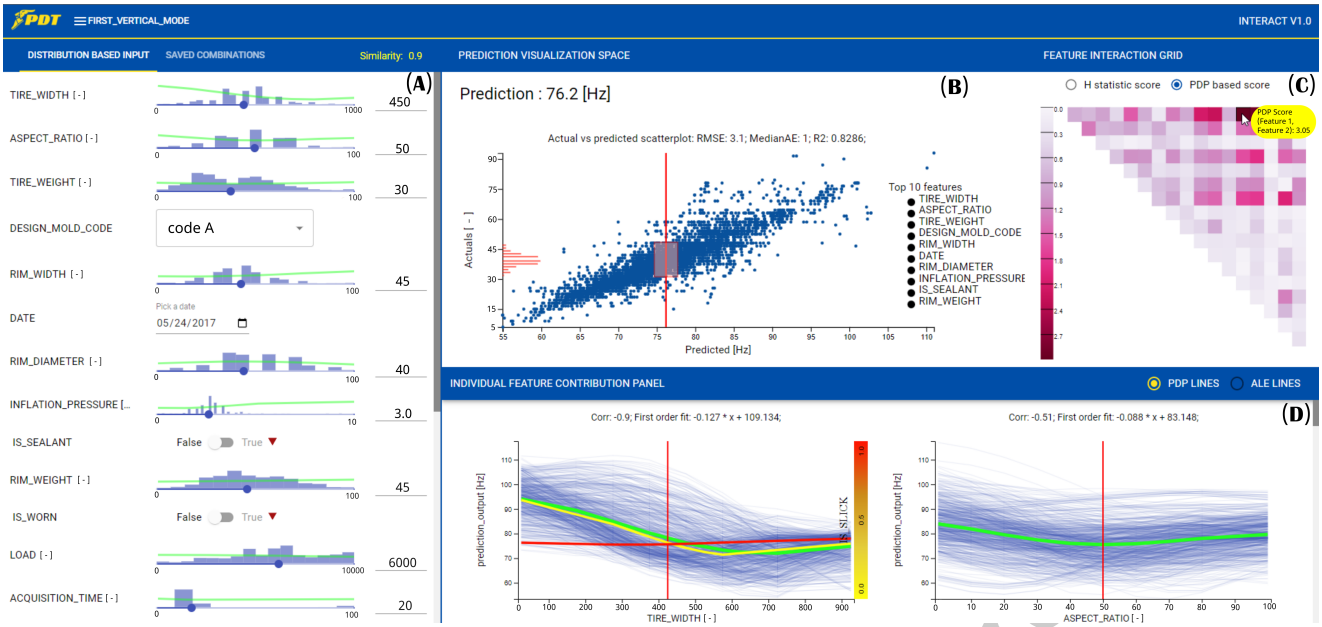

Figure 1. The graphical user interface of INTERACT showing the visualization for the first vertical resonance mode frequency[9] model. A) Data input panel: A set of scented widgets[10] for user control and display of the values of the most important model features. B) Actual vs. predicted plot: a scatterplot augmented with a confidence (red) box and an histogram of the actuals (plotted on the y-axis) for a given prediction. The legend shows the most important model features. C) Feature interaction matrix: Second degree feature interactions based on either the H-statistic or the Partial Dependence (PD) score. D) Individual Conditional Expectation (ICE) plots: a set of linecharts showing the individual contribution of each feature. The yellow to red curves show the interaction between the selected pair of features from (C).

pairwise interaction effects are also visualized using 2D contour maps[45,47] and 2D surface curves in a 3D-coordinate system[43,47]. The VINE[45] tool uses clustering in ICE/PDP line charts to visualize interaction effects in black-box models. Variable Interaction Networks (VIN) use small node-link network diagrams[48] to provide an overview of a dozen of pairwise interaction effects. Unlike most tools, INTERACT provides an interactive and scalable visual overview of all pairwise interaction effects and their relative strength.

What-if tools

Amer *et al.*[49] and Harries[50] describe the scenario use as "the creation of the description of alternative future realities". Golfarelli *et al.*[51] defines a scenario as a simulation of complex systems under a given hypothesis. What-if analyses are deemed useful in many domains, e.g., healthcare[52–54], social sciences[55–57], military[58,59], and were used heavily in the last couple of decades[60,61]. The most relevant to our work, visual what-if tools are *Prospector*[62], *The What-If Tool*[63] and *CoFFi*[64]. All systems use partial dependence (PD) concepts[65] to support overall and/or localized model inspection. While INTERACT proposes PD lines, we extend the explainability part with Individual Conditional Expectation (ICE) lines and Accumulated Local Effects (ALE) lines[65]. INTERACT differs also by a main focus on a fundamental new concept for what-if tools, which is the identification and support for understanding statistical interactions. Unlike previously cited tools, this aspect is required and useful in the context of virtual product design. Unlike other tools, INTERACT supports the comparative analysis of multiple alternatives (Figure 6).

Decision making

According to Edwards[66], the scope of the Decision Making (DM) theory is to predict, given two states A and B, which state will be chosen by an individual. Slovic *et al.* studied DM with respect to the decision environment, the different theories of DM and information processing in DM[67]. Other fundamental work concerns DM and cognition[68–71]. Milkman *et al.*[72] stress that in the DM context, errors are costly and will get even costlier, hence the need to improve DM processes. It is believed that AI can improve human analytical skills, DM abilities and creativity[73]. Along these lines, Bastani *et al.*[74] devised an ML algorithm that helps human users in DM tasks by using tips. Khosravi *et al.*[75] focus on combining DM and ML for flood susceptibility modeling. Healthcare is also a key area for improving DM with AI[76–79]. Duan *et al.*[80] give an overview of AI for DM and propose twelve directions for future research in this area. Some of these directions are further enhanced by Dwivedi *et al.*[81]. In both papers, the AI work is divided into supporting, augmenting, replacing or automating human tasks. INTERACT is designed to support and augment engineers' analytical abilities in their DM processes.

Tasks, data and users

Our research is based on frequent exchanges with domain experts, over a period of two years, leading us to address the formalization of user tasks as a primary step of our design. Also, two of the authors are domain experts (8 and 25 years in the field) and act as a "liaison", a role highlighted by Simon *et al.*[82], for a better definition of the tasks. The overarching goal of our target users is to design a product, i.e., a tire that meets a specific requirement, for example

a noise target. To do so, they need to factually decide on multiple characteristics governed by complex relationships.

Following Munzner's visualization design methodology[83], we provide below a characterization of data and tasks.

Data and users

Our work was conducted in collaboration with domain experts in the tire industry. Throughout the paper, we will illustrate our concept with the application of INTERACT to support acoustics engineers in the tire industry. They mainly ensure that the tire design meets all requirements in terms of acoustic comfort. A typical example of an acoustic engineer's work is to analyze why an existing tire does not meet the acoustic requirements and give tire design recommendations to reach the customer's acoustic target[84]. Among many possible recommendations, the engineer might propose to change the materials for a set of tire components, e.g., the material of the tread band, or to optimize the tread pattern design. Only two people among our domain experts have created ML models in the past. For example, the predictive model used in the case study (see page 8) was created by a senior engineer from the acoustics team.

The data pipeline at hand starts from audio recordings from various noise acquisition systems. These files are processed by the engineers to extract specific acoustic metrics quantifying acoustic comfort. These indicators are stored in relational databases along with tire design features and will be the focus of our study. The data used in our work is therefore tabular. The models are essentially serialized Python objects, but INTERACT can broadly query any regression model callable from a command prompt.

Tasks

In this section, we present the Domain Specific Tasks (DST) supported by INTERACT. The tasks are inspired by existing task categorizations concerning predictive visual analytics[85], trust in ML[25], visualization for ML[86]. We retained four high-level tasks from the literature and propose a new one related to the inspection of statistical interactions (DST2).

DST1: Visualize the most important features. While senior engineers might know which features are the most important and which "knobs" to turn first, a novice engineer needs support in the more subtle parts of the design process.

DST2: Detect and analyze statistical interactions. Statistical interactions complexify the iterative process of reaching a target. Even senior engineers need assistance when interactions arise. The support for identifying and understanding how features interact in a statistical sense (see definitions on page 2) is a required functionality.

DST3: Assess trust in predictions and model as a whole. Firstly, the domain experts need to know how good the underlying model is. General goodness-of-fit statistics must be available. Secondly, the users need to know if an input they use for model probing is in the training data. If not, there is less evidence supporting the prediction. Uncertainty presentation and cognitive load are also factors that highly infuence trust and which need to be considered when designing the system[87].

DST4: Create what-if scenarios. In the tire development context, building a physical tire is very costly and time consuming. A virtual tire development process allows many operations to be done *in silico*. INTERACT needs to support the creation of multiple what-if scenarios to reduce prototype manufacturing and testing cost and duration. For our domain experts and field, this includes the need to compare two or more virtual tires and make a factual decision on this basis.

DST5: Externalize knowledge through model explainability. A complex model can capture interesting and unknown forward paths that the engineer might explore to reach her targets. Revealing the potential effect of certain actions on reaching the target, like changing a specific tire design element, is also of high interest.

System description

In this section, we describe the coordinated multiple views of the INTERACT application (Figure 1). In line with the third level of Munzner's nested model[83], we motivate the proposed visual encodings and interactions supporting the tasks listed above, and discuss alternative visual designs. We also highlight the current practices of the target population of tire engineers. A demo of the application with non-confidential data is available at: https://viana.list.lu/interact.

Input space visualization

In **DST4**, the application domain expert needs to probe the model with different candidate designs, by tuning any available input parameter. We offer scented widgets: sliders, dropdown menus, toggles, and calendar widgets to set numerical, categorical, binary and date values respectively (Figure 2). Scenting[10] the sliders was crucial, as data is mostly numerical in our use cases. Often, tire engineers run exploratory data analyses in commercial software like Minitab[88] and SAS JMP[89]. They usually plot data distributions as bar charts, a popular choice in many areas[90]. Thus, INTERACT uses scented sliders showing the data distribution of the training data as a bar chart. Instead of using an arbitrary number of equally spaced bins, INTERACT uses a Numpy function to compute the best binning for each feature. We tested density plots too. The interpolated segments of the line charts created spurious rise and fall patterns, when a feature had gaps in its domain.

Towards **DST1**, the sliders are sorted top-down by feature importance in the ML model. The top features are indeed the ones that affect the predicted outcome the most; varying their values may help in appraising the model. INTERACT supports any regression ML model as long as feature importance values are provided.

For **DST5**, the sliders carry a second scent as a (green) line chart on top of the bar chart showing the expected prediction trend. At the user's discretion, the trend is computed using either a partial dependence plot (PDP) or an accumulated local effect (ALE) line[65]. These line charts share the same scale across all sliders to ease the comparison of the impact of features on the prediction. For boolean features, the toggle is scented with feedforward information[91]: a red or blue triangle shows the expected direction of change of the prediction, should the toggle be flipped.

In the top right corner of Figure 2, we display a score between 0 and 1 measuring the similarity of the current set of inputs to the training data based on a kernel method[92].

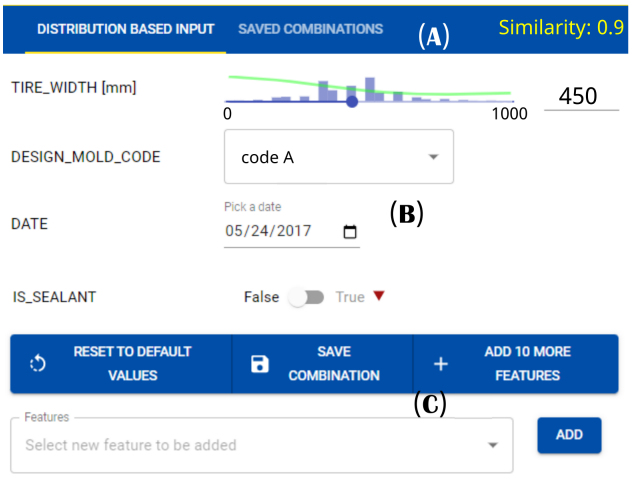

Figure 2. Control pane. **A:** Two tabs to show either the inputs panel or the comparative analysis panel. The similarity score of the current set of inputs compared to training data is shown in the top-right corner. **B:** Scented widgets[10] for various data types, sliders, dropdown menus, calendar widgets and toggles. **C:** Widgets to reset, save combination for comparative analysis or add features.

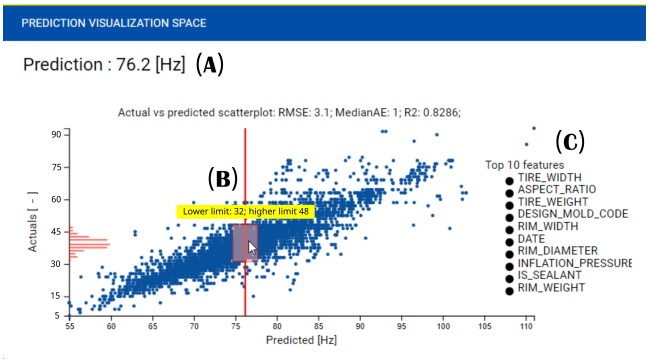

Figure 3. Prediction panel matching Figure 2. **A:** Prediction for current slider combination. **B:** Actual vs. predicted plot with linked cursor to current prediction. **C:** Top 10 most important features of the model.

The lower the score, the more cautious should the expert be regarding the prediction.

When the user finds a combination of input parameters yielding an interesting outcome, she may add it to the saved combinations for future reuse, e.g. comparative analysis. To support comparative analyses, the "saved combinations" tab (Figure 6) is laid out according to an augmented juxtaposition strategy. Techniques part of superposition or explicit encodings' families were excluded due to their known drawbacks in terms of visual clutter and decontextualization[93]. The user must be able to easily identify an alternative (virtual tire/product), and the latter needs to use the same units and the same scales across all dimensions as the original product (for instance, a control tire).

Another argument in favor of juxtaposition is the common practice and familiarity of our users with such designs within the organization. We reached our final design (Figure 6) after two iterations, while optimizing the layout and the scents. We use the percentage of change of an input as a scent, depicted by a triangle every 25%. The direction of change is encoded by color and by the direction of the triangles. The comparative panel can be set to only show the modified inputs across all alternatives. Knowledge externalization (**DST5**) can occur through questions like: "Which alternative yields the best prediction?", or "How do the inputs differ across the saved combinations?"

Prediction visualization

In **DST3**, the goal is to build trust in individual predictions and in the whole model. The trust we want to build with our system, in **DST3**, corresponds to the first, fourth and the fifth levels of trust of Chatzimparmpas *et al.*[25], concerning the raw data, the concrete model and the evaluation/user expectation.

In our context, the domain experts need to appraise the model for their decision-making process. Statistical software like Minitab[88] and SAS JMP[89] use common metrics like RMSE, R2 and MAE, and actual vs. predicted scatterplots to assess goodness of fit. In terms of alternate visual design, we wondered whether to plot predictions or actuals on the Y-axis of the scatterplot. It turns out that plotting the actuals on the Y-axis leads to a more correct model evaluation[94]. Hence, we use this design in INTERACT as in Figure 3, besides the popularity and suitability of scatterplots for regression models. In INTERACT, the prediction is displayed in textual form above the scatterplot and positioned as a red cursor (vertical line) on the plot. The prediction results from the current combination of inputs in Figure 2. Below the prediction, we display standard goodness-of-fit metrics, which provides the users with a way to gauge their confidence in the model. The legend reminds the top 10 most important features.

To trust a given prediction, the expert needs to situate it in the point cloud. When it falls in a dense area of the cloud near the identity line[95] the prediction is fairly reliable. If not, some caution may be needed. This calls for a confidence interval around the prediction[96]. To help the user assess such a confidence interval, INTERACT displays a red box enclosing 95% (box height) of the actual data around the prediction in the actual vs. predicted plot (Figure 3). A tooltip shows the lower and upper bounds of the box. In addition, a histogram alongside the Y-axis (in red) shows the distribution of the actual data (Figure 3). This provides even more context to assess the trustworthiness and reliability of the prediction.

Overview visualization of feature interactions

In **DST2**, domain experts are wary of statistical interactions between model features. The response of the model to an input feature often depends on the value of another feature. Ignoring feature interactions would instigate a tedious trial and error strategy to reach a target, without understanding how features interact. Current practice consists in modeling with standard least squares fits[97] in JMP or Minitab. The user must specify manually which interactions, if any, to include in the model fit[98]. In contrast, INTERACT computes automatically all pairwise interactions using either the H-statistic[43] or the PD-based metric[44].

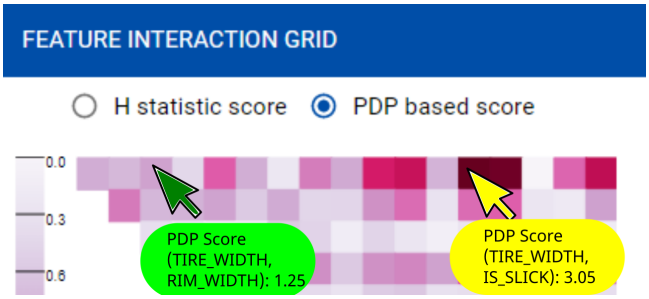

Figure 4. First four rows of the Interaction matrix. The yellow cursor points to the strongest interaction. The name of the features and the interaction score are displayed.

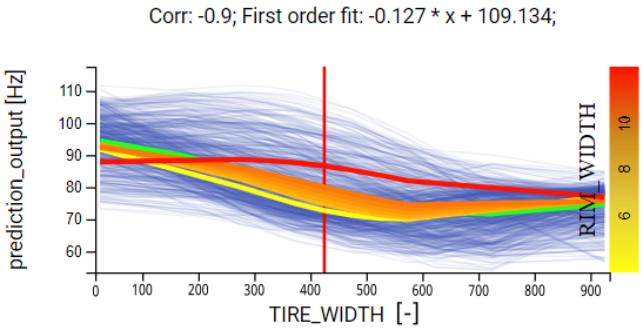

Figure 5. Individual feature contribution panel. ICE lines in blue. PDP in green. A click by the green cursor in Figure 4 populates the plot with second order interaction visualizations - yellow/red lines.

Since pairwise feature interactions may be modeled as a weighted complete graph, possible visual designs include node-link diagrams[48] and matrix visualizations. By applying a threshold on edge weights, we generated node-link diagrams, which were yet too cluttered. Finally, we preferred matrix visualizations, as they are better suited for link lookup in dense graphs[99]. For example, Figure 4 shows the first four rows of the interaction matrix for a modelled dataset of the first vertical mode of a tire[9]. The darker the color, the stronger the interaction between the pair of features. The user can quickly spot the highest interactions and get details on demand by hovering over a cell. This visual design scales up to dozens of features on a regular desktop screen.

Individual feature contribution visualization

In **DST1** and **DST5**, domain experts need to understand the sensitivity of the model to each input feature. They often use software to generate X vs. Y scatterplots to fit linear regression models between the predicted outcome and input features, along with statistical significance tests[100]. Even with a few features, the analysis of the resulting matrix of scatter plots is challenging[101]. Another practical limitation results from the fact that the observed data is usually much sparser than all possible combinations of a feature set, which makes it difficult to clearly isolate the individual contribution of a feature of interest. An alternate method that allows the user to rapidly inspect the relation between a feature of interest and a target is the partial dependence (PD) plot (Figure 5). The equation corresponding to the first order fit of the PD line and the Pearson's correlation coefficient[102] are shown above the chart. In the presence of correlated features, the Accumulated Local Effects (ALE) can be used instead[65]. INTERACT provides both PD and ALE line charts to give an overview of the full data set.

Beyond global methods, finer phenomena can be inspected using the ICE lines (blue lines in Figure 5). For each point in the dataset, a line is created by varying the independent feature stepwise and fixing the values of all other features.

Figure 1 D shows two ICE plots with the Partial Dependence (PD) line in bold green. The left plot presents a general descending trend (PD) for `TIRE_WIDTH` (green PD line). The ICE lines (in blue) tend to agree with this trend. This behavior is expected by experts and agrees with recent literature[103]. The wider the tire, the heavier the tread and the lower the first vertical resonance frequency. The vertical spread of the ICE lines also gives an idea of the range in which the prediction may lie with respect to the feature on the X-axis. Tightly bundled ICE lines point to the lack of interacting features and to a low influence of other features on the prediction. As ICE lines spread out and form crossing patterns, feature interactions are more likely[65].

User Interaction

INTERACT is a coordinated multiple view system, designed to support what-if analyses built on top of a predictive ML model. More than a mere collage of the individual views described so far, the linked views support a cohesive analytic workflow driven by user interactions.

What-if query specification. Starting from Figure 1 A, the user can express a what-if query by pulling a few sliders to set the values of selected input features. The similarity score in the top right corner of the data input panel (Figure 1 A) shows how close the new input configuration is to the training data. The underlying ML model is queried for a prediction. Figure 1 B is updated to reflect the new prediction textually at the top and graphically, by adjusting the position of the red cursor, the related confidence box and histogram. The values of the input features are also cascaded to any related charts, e.g., the red cursor in each ICE line chart in Figure 1 D. This helps the user to locate the prediction in the feature space. While only the top most important features are visible by default in Figure 1 A, the user can interactively display more features (Figure 2 C). The interaction heatmap in Figure 1 C is expanded accordingly to include additional rows and columns. Likewise, the related ICE line charts are added (Figure 1 D). Since what-if analyses are often based on trial and error, INTERACT provides a save and reload functionality. The final decision making usually considers multiple saved candidates. We offer a simple yet effective visualization to compare the chosen alternatives.

Figure 6 shows three candidate tire designs, yielding the same prediction value (76.2). The user can decide which alternative suits best her needs, based on other, unmodeled criteria such as cost or manufacturing constraints.

Drilling down into statistical interactions. Continuing the analysis in Figure 1 C, the user's attention is drawn to a few dark cells in the heatmap corresponding to pairs of strongly interacting features. By clicking on any cell, the individual feature contribution visualization in Figure 1 D is overlaid with a set of new lines, colored using a yellow-to-red

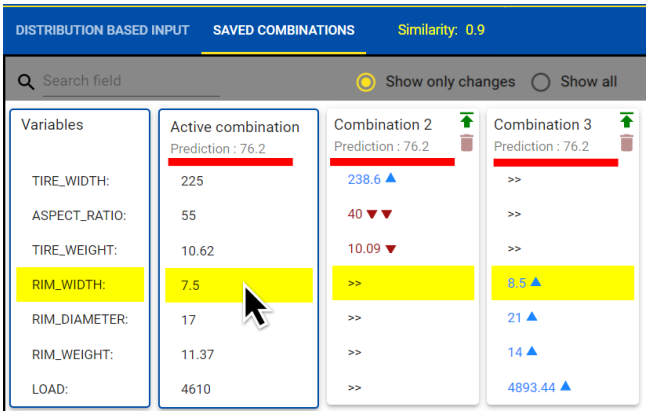

Figure 6. Three combinations showing the same prediction level (above red line) but with different solutions. The reference is the active combination (first column); the scents (triangles) and equality (>>) are with respect to the active combination.

gradient. While the X-axis of the line chart corresponds to the selected row in the heatmap, the color encodes the level of the second feature, i.e., the column in the heatmap. We chose to display feature names on demand only when more than five features are shown in the heatmap, to avoid label clutter. We also do so because feature names tend to be very long in our context of use, and truncating them is not helpful. The color scale is shown on the right side of the plot. These lines correspond to the second order partial dependence lines according to a predefined grid, e.g., a decile-based grid. The user may hence inspect the trends of these lines and identify the regions where statistical interactions occur.

Technical implementation

INTERACT is implemented following a state-of-the-art Web architecture with a FastAPI server at the backend, a JavaScript middleware using Node.js, and a JavaScript front-end using React, Redux and D3. To ensure high availability of the application, the complete architecture is containerized using Docker[104] and orchestrated using Kubernetes and Terraform in a cloud environment.

Evaluation

Our evaluation methodology was inspired from existing efforts in categorizing and guiding evaluation in visualization research[83,105–111]. Referring to Lam et al.[105], we evaluated INTERACT from two perspectives, namely Visual Data Analysis and Reasoning (VDAR) and User Experience (UE). We conducted a qualitative user study with a group of acoustic engineers in the tire industry (see User study) and a case study with two domain experts (see Case study). The evaluation procedure and hypotheses were preregistered at https://osf.io/-clickable. We also emphasize that user study participants and experts involved in the case study do not include any of this paper's authors.

User study

Methodology. We ran a first pilot study with 11 engineers, aiming to assess the usability of INTERACT. The first pilot study was also registered at https://osf.io/-clickable.

This study led to a small-scale deployment of INTERACT within the Goodyear organization, and many suggestions for improvement.

Six months later, we ran a second pilot study, preregistered at https://osf.io-clickable, to evaluate the improved software in a real use context, i.e., with tire engineers and tire related data. We started with a training session aiming to introduce the main functionalities of INTERACT, train on solving tasks, and get familiar with the tool through concrete examples. Next, we organized the user study. User data and feedback were collected through: 1) a survey of demographic and occupational data, 2) a semi-structured focus group and, 3) an anonymous exit questionnaire. The training and user study took two hours in total. The results of this second pilot study are discussed below.

Participants. The training was organized with ten engineers, all of whom were domain experts in tire acoustics from the Goodyear organization. The median experience was 13 years with an inter-quartile range of 10 years. Two of them have previously created ML models and two more have participated in ML model validation. The remaining six have no experience with ML. All participants were co-located and used commodity 17" laptops. More demographic information is provided in the supplemental material.

Focus Group. After the training, the domain experts took part in a semi-structured focus group[112], in the form of an open discussion around prepared questions in relation to our research objectives and domain-specific tasks. One of the co-authors facilitated the discussion while another one took notes. We also recorded participant voices, later transcribed and analyzed question by question. Each answer was split into one or multiple semantic units (one idea per unit text)[113]. Following a 'peer debriefing' method[114], three independent coders labeled these units, and then met and agreed on the following set of tags: utility, UX, learnability, enjoyment, extensibility.

The **utility** tag was by far the most important tag in our coding effort. All comments were positive and highlighted how the features of the tool can support participants' daily work. The tool was deemed an asset for discussing what-if scenarios with internal and external customers, for detecting and analyzing statistical interactions, for knowledge externalization and model explainability.

In *what-if* discussions, the tool can be used to answer customers on the spot rather than postponing the analysis. *"It will help create a data-driven way of replying to customers"*. Besides providing a prediction, *"knowing the uncertainty is very valuable"*. This will further allow to build trust in predictions (DST3). In this regard, the PD line visualization also was appreciated, with a participant saying: *"Referring and following it would provide really good feedback in conversations with customers"*. Regarding the what-if capabilities, the users liked the real-time model probing and its smoothness (*"I like the speed of the tool"*). This capability was even more interesting for younger engineers, who say they *"don't have the knowledge base to fall back on"*. The visualization of statistical interactions was perceived as having a high added value as it allows going beyond guidelines. *"More information on a specific variable can be extracted if a guideline rule is wrong"*.

Knowledge externalization and model explainability *"helps from a fundamental understanding to better comprehend the sensitivities and which factors are most important"*. Other comments on this topic include: *"It provides a simple way to understand a model"* or *"Leveraging data that I never accessed expands my capabilities beyond my personal knowledge/experience"*.

The **UX** tag encompassed mainly request for improvements of the user interface and requests for new functionalities. The latter often concerned open research questions, which we discuss thoroughly in the next section. Some bugs were also reported and fixed soon after.

In terms of **learnability**, the users asked for more contextual help and more training material. This was particularly the case with novice engineers who were somewhat new to statistical interactions. The concept is a bit more complex to grasp initially. Yet, all participants deemed INTERACT *"intuitive after using it"*.

The **enjoyment** tag was mainly assigned to comments regarding the application as a whole. The users liked the visualizations, the real-time model probing, but also the easy identification of the most important statistical interactions.

Extensibility in terms of the number of models that can be added to INTERACT was raised. Pending a formal assessment, the tool can offer many more models, owing to its cloud computing architecture and a caching mechanism.

Exit questionnaire. The post training questionnaire took the form of a 7-level Likert scale survey [115] (see supplemental material): 1- Strongly disagree; 2- Disagree; 3- Somewhat disagree; 4- Neutral; 5- Somewhat agree; 6- Agree; 7- Strongly agree. Its creation and analysis followed the best practices [116–118]. Where suitable, the question replaced ("Agree") by ("Useful"). Each Likert scale was composed of several Likert items (sub-questions) that were closely related. We analyzed the results using the one-sample t-test [119] and its non-parametric equivalent, the one-sample sign test [120]. This is due to conflicting opinions regarding whether Likert scales should be considered as ordinal or numerical [116–118]. We applied both strategies, and in each case, INTERACT obtained statistically significant positive outcomes ($p < 0.05$) in support of our Domain Specific Tasks (see Tasks). The neutral point (mid-point of the Likert scale) served as the benchmark against which we calculated the p-value. Table 1 shows the results of the Likert scales only. More details about the individual Likert items are provided in supplemental material.

Case study

This case study aims to observe and report on how INTERACT can be used by domain experts to address real business needs. We ran this study with two senior experts in tire acoustics, none of whom is an author of this paper. We used the First Vertical Mode Model, shown in Figure 1 as background data and model. We structure the following account into two subsections arising from the actions performed by the two experts. The problems they wanted to solve with INTERACT were akin to their daily tasks and are representative of two main challenges acoustic engineers face in their work: 1) problem solving; 2) knowledge externalization and exploratory analysis.

Problem solving. Car manufacturers, provide tire performance specifications including a list of metrics to be met during the tire design phase, e.g., acoustics and mileage for a tire fitment on a specific vehicle. This is a common trait in all industries where users specify a set of desirable product characteristics. In this section, we will focus on a hypothetical yet realistic case where meeting the customer target requires, for instance, the tire's first vertical oscillation mode (natural frequency) to be below 70 Hz. To the layperson, we highlight that all actions presented below relate to influencing tire mass or stiffness. Classical mechanics show that the natural frequency in a mass-spring set up can be expressed in the form $\sqrt{\frac{k}{m}}$ with k being the stiffness and m the mass [121]. Increasing mass and decreasing the stiffness will lower the natural frequency. This is what our expert tried to achieve with INTERACT. To start with, the expert set the mandatory values provided by the customer for tire size, i.e., tire width, aspect ratio and rim diameter, using the input sliders, hence obtaining a prediction of 77 Hz in Figure 1 B (DST4). Next, he changed the rim width, a parameter which could be changed in a limited range. Looking at the PD line on top of the slider (5th input feature in Figure 1 A), reducing the value would also lower the first vertical mode frequency. Shrinking rim width from 7.5" to 7.0" led to a predicted 74.3 Hz, which is a small improvement. Likewise, the inflation pressure was reduced to 2.0 Bars to reach a mere 74.2 Hz. The next attempt consisted of adding a layer of sealant material, which happens to make the tire heavier [122] (first toggle in Figure 1 A). As shown by the red triangle next to the toggle, adding a sealant would get us closer to the target. This took us to 67.5 Hz, even better than the customer request. While this effect is known to decrease the first vertical mode frequency [103], we might prefer a less drastic design solution to achieve the same performance. He changed the toggle back to investigate other options. One of the top features in the list was tire weight. Despite a high ranking (3rd), the PD line was quite flat. Looking at ICE lines for this feature, typical crossing patterns pointed to statistical interactions (DST2). He then inspected the interaction matrix (Figure 1 C). The highest interaction was with seasonality (3rd row, before-last cell). Winter tires exhibited a falling trend, while summer tires were rather flat. Hence, he checked the current status. INTERACT predicted 74.2 Hz for summer tires. Toggling the season to winter took the prediction below 70 at 69.2 Hz. This could probably *"make sense"*, he said. Across the market, winter tires typically have more rubber, i.e., mass in the tread area (DST5). *"In addition, the ICE lines are useful to evaluate the feasible design space. We can look at the range (thickness of blue lines) of ICE lines and already estimate how challenging a target could be"*, the expert said. He then checked other statistical interactions, acknowledging interesting information to be investigated more in depth.

Knowledge externalization and exploratory analysis. Engineers in tire acoustics must give recommendations to reduce tire noise. They usually make suggestions based on company guidelines or simulation tools. Often, these guidelines can be generic and high-level. Other practices include the comparison with previous projects, experience sharing and potentially back-to-back analysis [123,124]. These

Likert scale \ Rating and statistics	D-- (1)	D- (2)	D (3)	N (4)	A (5)	A+ (6)	A++ (7)	Mean	Median	Stdev	IQR	Prob>t	Sign p val	N
The views/visualizations are useful	0.0	0.0	0.02	0.16	0.18	0.38	0.26	5.7	6-Agree	1.1	2	$<10^{-4}$	$<10^{-4}$	50
The application is useful	0.0	0.0	0.0	0.0	0.05	0.35	0.6	6.6	6-Agree	0.6	1	$<10^{-4}$	$<10^{-4}$	20
Identification of the most important features	0.0	0.04	0.0	0.16	0.14	0.36	0.3	5.7	6-Agree	1.3	2	$<10^{-4}$	$<10^{-4}$	50
Understanding of statistical interactions	0.0	0.1	0.0	0.1	0.2	0.3	0.3	5.5	6-Agree	1.5	2	$<10^{-4}$	$<10^{-4}$	20
Trust in model predictions	0.0	0.0	0.05	0.15	0.18	0.3	0.32	5.7	6-Agree	1.2	2	$<10^{-4}$	$<10^{-4}$	40
Interact supports what-if scenario creation	0.0	0.03	0.0	0.03	0.3	0.4	0.23	5.7	6-Agree	1.1	1.3	$<10^{-4}$	$<10^{-4}$	30
Interact allows to externalize knowledge	0.0	0.0	0.0	0.03	0.17	0.43	0.37	6.1	6-Agree	0.8	1	$<10^{-4}$	$<10^{-4}$	30

Table 1. Summary of the anonymous exit questionnaire. The columns comprise the Likert scale ratings: Strongly Disagree D--(1), Disagree D-(2), Somewhat Disagree D(3), Neutral N(4), Somewhat Agree A(5), Agree A+(6), Strongly Agree A++(7). Descriptive statistics are displayed in the following columns, including mean, median, standard deviation, interquartile range, one-sided t-test p-value (prob>t), sign test p-value, and the number of items answered for each corresponding Likert scale (sub-questions × number of participants). Heatmap annotations are expressed in percentage out of N. The descriptive statistics, such as the mean, are computed based on the rating values, i.e., Strongly disagree = 1 and Strongly agree = 7.

methods work but are time-consuming and add a data retrieval burden. Yet, drawing conclusions from few data points is more risky. Experts need a vast domain knowledge to answer such requests. Before using and trusting the model (DST3), both experts sought to verify that the model agrees with physics. They inspected and confirmed the most important features (DST1) but also audited the actual vs. predicted plot. Next, the PD lines (in green) were analyzed. The amplitude and direction of change with various features were inspected, looking for directly or inversely proportional relations. *"The green line is showing a trend which is already good, this is something I was looking for!"*, an expert said. The directions and amplitudes made sense and were confirmed for most features. Another observation was linked to the inspection of the ICE lines (Figure 1 D) for tire width. One can observe various patterns in the data. For instance, a group of lines has a falling trend, reaches a minimum then goes up. Others exhibit a plateau or a rising trend, then go down. *"This means that there are groups of tires that behave in some way and others differently. Based on hundreds of tires, it's really good!"*. To explain this difference, for instance in Figure 5, we can see that these two trends are reflected and are grouped by rim width.

One of the experts made another hypothesis linked to a temporal evolution of another noise metric. In essence, she thought that due to progress made in the past years, this metric increased until it reached a peak and then plateaued. This trend was easily confirmed in INTERACT (DST5).

Discussion and Conclusion

In the past decade, many visualization-assisted ML tools have been proposed. While many techniques and visualization tools can explain the learned model, or create what-if scenarios, to our knowledge, none supports the analysis of statistical interactions. INTERACT was proposed to fill this gap and was validated in the context of virtual tire design. INTERACT has new capabilities such as real-time, interactive what-if scenarios, statistical interaction analysis, comparative analysis for input-prediction pairs and means to assess the trustworthiness of predictions. Besides having two domain experts heavily involved in the tool design (the two co-authors), INTERACT was evaluated by ten independent domain experts who found it useful in their applied context and provided feedback for further improvements.

Limitations

Currently, INTERACT supports regression models only. So far, we developed scented widgets for numeric and binary input features. Categorical and date-type features are also supported by INTERACT, albeit with limited feedforward information[91]. The computation of interaction scores for categorical input variables has yet to be integrated in the tool. Also, an alternative to ICE plots may be needed for multilevel categorical features, as the levels might lack a natural order. The tool can also be extended to support classification models. Then, we could reuse some of the current visual designs, e.g., the input panel (Figure 2), but the other views may have to be redesigned.

INTERACT currently approaches the trust perspective from three key angles: examining the raw data, assessing the concrete model, and understanding user expectations (refer to TL1,4,5 in *Chatzimparmpas et al.*[25]). In this paragraph, we will outline the support provided in INTERACT to build trust in the model as well as possible extensions that might be required in other contexts. For raw data analysis (TL1), we employ input feature distribution visualizations, enabling us to visualize the data used to train the model and determine whether its ranges align with the problem at hand. Going beyond what is offered in INTERACT, more detailed information such as point by point data inspection and data attrributes descriptions could also be useful.

To evaluate the model's quality, INTERACT presents standard goodness-of-fit metrics. These metrics can help the user to build trust in the model across the design space (TL4). Going beyond building trust in a model, it might be desirable to be able to compare multiple versions of a model, for example when retraining the model as new data becomes available, as mentioned in TL4.

Furthermore, we offer trust-building visualizations related to the model, such as the PDP, ICE and ALE plots. The latter allows users to quickly observe whether outcomes vary as expected with specific inputs - TL5 (see Input space visualization). Besides fostering trust in the model as a whole, we provide localized goodness-of-fit information, as illustrated in Figure 3. This information is generated based on the distribution of actual data in the vicinity of the current prediction. We supply a prediction interval that encompasses 95% of the actual data. However, it's worth noting that this interval isn't solely based on the current prediction but rather on all predictions within its neighborhood. To further enhance the accuracy of individual predictions and

reduce uncertainty, techniques like Gaussian processes can be employed[125]. Addressing this specific limitation is a part of our future development plans. Similarly, more work for the development of visualization of causality[126] can be undertaken to further enhance trust.

The feature interaction metrics used by INTERACT can scale up to higher-order interactions, but the current version of the tool displays second-order (i.e., pairwise) interactions only. The interpretation of second-order statistical interactions is known to be hard, and things worsen in the case of higher-order interactions[47,127]. This is why we chose not to go beyond the second order. As the interpretation of interactions relies also on domain expertise, INTERACT puts the expert in the loop by affording her to visualize interaction effects and anticipate the direction and magnitude of their impact on predictions. The ease of use and the snappy response achieved by INTERACT enables the expert to examine thoroughly and build trust in the ML model eventually.

Instead of shipping a closed list of pretrained ML models, some of our study participants wanted to use the tool with improvised models. Similarly, users might prefer a different arrangement of the sliders, which are currently sorted based on model's feature importance metrics. Users might be accustomed and/or grouped to some preferred order and this poses visual search challenges. We will improve the flexibility of INTERACT in these directions in the future.

Some users also wanted a more detailed instance-based view of the data. The tool currently allows to inspect individual data points only by mousing over the line plots or scatter plots.

Applicability beyond tire design

INTERACT is a visual what-if tool based on regression models with enhanced support for the analysis of statistical interactions. The tool was developed in collaboration with tire industry engineers. The validation of the tool was mainly based on a user study and a case study in this industry, presented earlier in this paper.

Nevertheless, it is reasonable to believe that our approach is transferable to other application areas where regression models are used to make predictions based on multiple inputs. This is justified by the fact that four out of five tasks (DST) that we support with INTERACT are well-documented in the XAI litterature (see Tasks). This paper puts forth one new task, which is the analysis of statistical interactions (DST2). We believe this is an underexplored area as, to our knowledge, no other what-if tools propose such analysis (see Related Work). We propose a solution that fits the tire industry needs and call for more work in this direction. Although we cannot assert the relevance of statistical interactions in general for every regression problem, we strongly believe it is applicable within the broader scope of product design, where physics laws often induce such feature interactions. In this context, models can be tested against the physics of objects, requiring from experts to assess the behavior of models and hence their trustworthiness and reliability. A number of questions may guide the transfer of our approach to other fields: 1) Are higher-order interactions important to the target fields? If so, we still miss a way of visualizing them. 2) Are statistical interactions for categorical features essential to the target fields? If so, how can they be visualized? 3) Do other fields need to combine statistical interactions with optimization? Answering positively to any of these questions is grounds to not use INTERACT in its current form.

Research opportunities

We base our thoughts for future work on our experience while developing INTERACT, but also on the expert feedback received at various occasions, such as the focus group.

Single vs. multi-model approaches. INTERACT activates and queries one model at a time currently. It allows users to understand the active ML model and use it, e.g., for product design. Yet, in applied environments, experts usually need to meet many, often antagonistic constraints, which may take advantage of multiple models activated in parallel. An example of such work is VisProm[128]. In the future, we would like to test this system in our applied context, i.e., virtual tire development. We think more research is needed to: 1) explore the suitable design space of visualizations for multidimensional inputs together with multi-model outputs; 2) provide useful feedforward information in such contexts; 3) scale the visualizations of statistical interactions up for multi-model approaches.

The field of **Multi-Criteria Decision Making (MCDM)** is very rich (see Related Work). Many MCDM techniques were devised to help domain experts. We believe that this line of work can be even more useful, for instance, if combined with state-of-the-art multi-model XAI visualization tools. MCDM is by default multi-model and multidimensional and fits many problem-solving techniques encountered in product design. We call for more visualization work integrating MCDM with multi-model explainable ML approaches in various fields, not limited to product design.

Provenance and storytelling. When using what-if tools, users typically aim to reach one or a few working solutions to their problem. Often, beyond the interaction with the system, these outcomes are discussed, in other contexts with other colleagues or with management representatives. To mantain a consistent and structured dialogue, information related to the provenance[129] of the solutions is beneficial. The presentation of these solutions as a narrative storytelling[130] can further increase the cohesion of the solutions with the overarching goal. Therefore, we call for more work in integrating provenance analysis and storytelling with what-if tools.

Feature importance in ML methods is defined differently across model types, which makes it hard to understand or compare feature importance metrics[131]. Researchers might invent ways to design visualizations to allow the user to construct and visualize her own feature importance metric.

Dimensionality reduction (DR). INTERACT currently provides ICE lines for the analysis of individual features. Recent work has proposed to use clustering to aggregate these lines[45]. We wish to understand how DR techniques could be combined from a visualization perspective with supervised learning. Is DR useful in what-if analyses?

Lessons learned for what-if tools

Reflecting on 30 months of development of INTERACT, the following takeaways may interest visualization researchers.

Instant feedback. Systems such as INTERACT need to provide instant feedback. When multiple scenarios need to be assessed rapidly, the expert cannot wait long to obtain a prediction. In contrast to the "What-If Tool"[63], INTERACT computes predictions on-the-fly as the user moves input sliders, without waiting for the user to hit a "run" button. Users get, hence, a smoother experience.

Explainability mechanism. When exploiting complex models, their outcomes might be unpredictable or conflicting with expected behavior. What-if tools need to support the end users in understanding how the application reached its output. This will further enhance their trust in the system and allow for constructive criticism, for example for improving the underlying models. In our case, besides the classical ICE, ALE and Partial Dependence visualizations, we use the concept of statistical interactions to support this goal.

Comparative analysis. The ability to create and save multiple scenarios is essential in the context of what-if tools. The users need to be able to effectively visualize and reload previous states effortlessly during a working session and beyond. In the context of product design, we relate such a need to the creation of several virtual designs before manufacturing. Iterating from a previously saved virtual design and comparing it to the current design is a standard need in such contexts. The selection of a saved design among multiple prior designs needs to be affordable. The user needs to be supported in the choice of one of the saved alternatives for instance by having some visualization that helps in assesing the related model prediction or its configuration compared to others. We could even argue that the history or provenance information of the saved virtual designs could be of interest.

Feedforward information. The simplest what-if tool could probably allow for the creation of multiple scenarios by trial-and-error. We argue that this method would be too tedious in the context of product design and could lead to a loss of user engagement. We believe that the user needs to be accompanied in their search for a solution. We use and recommend the use of scented widgets displaying feedforward information to guide the user in the design space. Visual cues need to be available and affordable to the user for reaching their final goal and anticipate on actions.

Automation and need. Another consideration that needs to be taken into account when creating what-if tools is whether there is a real need for such a tool. We think typically of situations where optimization engines could provide successfully the expected results and the need for a what-if tool is not fully justified. Like data visualization is not always the most-suited solution for data-centric problems, a what-if tool is suitable when an algorithm (e.g. optimization algorithm) would require more information than is available to solve the problem, and/or the task is not fully clear (see the information location and task clarity dimensions described by Sedlmair et al.[132]). However, users need to benefit of as much automation as possible to reach their goals. They should focus on their decision making process without being distracted by usability issues. Nonetheless, automation must be carefully designed to not alter the exploratory nature of the what-if tool.

Realism. Often, models will provide a prediction even if the inputs do not make sense in reality. For example, a model that predicts maximum speed of a car based on a number of inputs such as car brand, engine type and, year of production, could accept an input such as a V8 diesel engine from 1992 on an electric car. While such an input is semantically void, the model will still provide a prediction. The industry is gradually adopting virtual design to reduce waste and needs to provide new and virtual inputs to models which can fairly interpolate/extrapolate. Users need an indication of the realism of such an input in light of the existing data/reality, or even the ability to compare and interactively select the most similar, yet realistic data points. Today, INTERACT uses a similarity metric (with respect to training data) which puts equal weights on the full set of inputs. Given the high dimensionality of the data, the resulting similarity scores lack sensitivity. Future research might focus on assigning weights to features, like in WeightLifter[133], and guiding the user towards feasible design spaces, similar to the TOP-slider[134]. This was a strong feedback from the focus group that we would like to tackle in the future.

Maintainability of what-if tools. What-if tools use models behind the scene. While model-agnostic approaches provide a good separation of concerns between the visual user interface and the underlying models, an obstacle still exists in the way of long-term deployment of what-if tools. Indeed, such deployment raises classic software architecture considerations in terms of versionning and compatibility issues for both the what-if tool and the models. An even better separation of concerns and maintainability could be achieved through a miscroservices architecture[135], i.e. use individual APIs for each of its models to avoid such issues in future.

Interoperability. Lastly, what-if tools need to be integrated in broader ecosystems where other software solutions are used by domain experts. What-if tools need to be interoperable, for example by their ability to load and export common data formats. Users could, for instance, extract data from the system and use it in their own presentations and storytelling.

General lessons learned

Visualization guidelines and preregistration. When designing visualization systems together with domain experts, the visualization researcher needs to exploit existing practical guidance. Throughout our project, we proactively recognized and confirmed most of Sedlmair et al.'s pitfalls[132], which helped greatly in our collaboration and developments. In addition, we found it helpful to preregister our evaluation/validation protocols. This allowed us to be better prepared and to spot potential methodological flaws early on.

Pre-existing software environment. When creating a new visualization system, the visualization researcher should analyze the software already in use by the target population, including commercial applications. This simplifies the discussion with domain experts, as they often refer to

visualizations and methods available in this type of software. This will also facilitate discussions about change management with the target users. The risk of having a comment such as "but this already exists in software X" is hence mitigated. An example of this in our case was the JMP statistical software[89]. This might also help to delineate the new contribution. Knowledge about company software development practices is also important to facilitate the deployment of the new application, e.g., to ensure its compatibility with existing orchestration tools.

Be or work with a liaison. The design of effective visualizations relies on a sufficient knowledge of the field for which the software is intended. When the visualization researchers have no background in the target field, it's crucial to be immersed in the domain and work with field experts. Sitting in a few meetings to gather requirements is not enough. Such collaboration must result in a team member taking the role of the liaison[82], i.e., a person with sufficient knowledge of both the target field and the visualization field to facilitate the interdisciplinary communication required for a successful project.

Acknowledgements

We would like to thank Prof. Seppe vanden Broucke from KU Leuven for authorizing the use of his H-statistic code[136], and Alexandre Dillon for GUI improvements. We thank Julien Vaissaud for the fruitful collaboration on integrating acoustic ML models and his great feedback. We thank Dr. Peter Kindt and Dr. Doris Maus for their precious comments as domain experts and their availability for the case study.

Funding

This work was supported by the Luxembourg National Research Fund (FNR) grant #14221651 and the Goodyear Innovation Center Luxembourg (GICL).

ORCID iDs

Vasile Ciorna https://orcid.org/0000-0003-2380-1655
Guy Melançon https://orcid.org/0000-0003-3193-7261
Frank Petry https://orcid.org/0000-0002-7951-8712
Mohammad Ghoniem https://orcid.org/0000-0001-6745-3651

References

1. Kwon BC, Choi MJ, Kim JT et al. RetainVis: Visual Analytics with Interpretable and Interactive Recurrent Neural Networks on Electronic Medical Records. *IEEE Transactions on Visualization and Computer Graphics* 2019; DOI:10.1109/TVCG.2018.2865027. 1805.10724.
2. Gotz D and Borland D. Data-driven healthcare: challenges and opportunities for interactive visualization. *IEEE Computer Graphics and Applications* 2016; 36(3): 90–96. DOI:10.1109/MCG.2016.59.
3. Stirnberg R, Cermak J, Kotthaus S et al. Meteorology-driven variability of air pollution (pm 1) revealed with explainable machine learning. *Atmospheric Chemistry and Physics* 2021; 21(5): 3919–3948. DOI:10.5194/acp-21-3919-2021.
4. Rautenhaus M, Böttinger M, Siemen S et al. Visualization in meteorology—a survey of techniques and tools for data analysis tasks. *IEEE Transactions on Visualization and Computer Graphics* 2017; 24(12): 3268–3296. DOI:10.1109/TVCG.2017.2779501.
5. Sukhija N, Tatineni M, Brown N et al. Topic modeling and visualization for big data in social sciences. In *2016 Int. IEEE Conf. on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress*. IEEE, pp. 1198–1205. DOI:10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.0183.
6. Gunning D, Stefik M, Choi J et al. XAI-Explainable artificial intelligence. *Science Robotics* 2019; 4(37): eaay7120. DOI:10.1126/scirobotics.aay7120.
7. Chatzimparmpas A, Martins RM, Jusufi I et al. A survey of surveys on the use of visualization for interpreting machine learning models. *Inf Vis* 2020; 19(3). DOI:10.1177/1473871620904671.
8. Gent AN and Walter JD. In *Pneumatic tire*. 2006.
9. Kindt P, Sas P and Desmet W. Measurement and analysis of rolling tire vibrations. *Optics and Lasers in Engineering* 2009; 47(3-4): 443–453. DOI:10.1016/j.optlaseng.2008.06.017.
10. Willett W, Heer J and Agrawala M. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Trans Vis Comput Graph* 2007; 13(6): 1129–1136. DOI:10.1109/TVCG.2007.70589.
11. Gillespie TD. Fundamentals of vehicle dynamics. Technical report, SAE Technical Paper, 1992. URL https://www.sae.org/publications/books/content/r-114/.
12. Wang X (ed.) *Automotive tire noise and vibrations: Analysis, measurement and simulation*. Butterworth-Heinemann, 2020. ISBN 9780128184103. DOI:10.1016/C2018-0-02431-7.
13. Clark S and Dodge R. *The Rolling Resistance of Pneumatic Tires. Final Report*. Report, U.S. Department of Transportation, NHTSA, Office of Research and Development, 1979.
14. Jazar RN. *Vehicle dynamics*, volume 1. Springer, 2008.
15. Pierce L. *Acoustics*, volume 3. Springer, 2019.
16. Wiener FM. Experimental study of the airborne noise generated by passenger automobile tires. *Noise Control* 1960; 6(4): 13–16. DOI:10.1121/1.2369419.
17. Lee J and Ni A. Structure-borne tire noise statistical energy analysis model. *Tire Science and Technology* 1997; 25(3): 177–186. DOI:10.2346/1.2137539.
18. Hastie T, Tibshirani R, Friedman JH et al. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009. ISBN 978-0-387-21606-5.
19. Iwao K and Yamazaki I. A study on the mechanism of tire/road noise. *JSAE review* 1996; 17(2): 139–144. DOI:10.1016/0389-4304(95)00004-6.
20. Guidotti R, Monreale A, Ruggieri S et al. A survey of methods for explaining black box models. *ACM computing surveys* 2018; 51(5): 1–42. DOI:10.1145/3236009.
21. Endert A, Ribarsky W, Turkay C et al. The State of the Art in Integrating Machine Learning into Visual Analytics. *Computer Graphics Forum* 2017; 36(8). DOI:10.1111/cgf.13092.
22. Hohman F, Kahng M, Pienta R et al. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Trans Vis Comput Graph* 2019; DOI:10.1109/TVCG.2018.2843369. 1801.06889.

23. Yuan J, Chen C, Yang W et al. A survey of visual analytics techniques for machine learning. *Computational Visual Media* 2021; 7(1): 3–36. DOI:10.1007/s41095-020-0191-7.
24. Arrieta AB, Díaz-Rodríguez N, Del Ser J et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 2020; 58: 82–115. DOI:10.1016/j.inffus.2019.12.012.
25. Chatzimparmpas A, Martins RM, Jusufi I et al. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum* 2020; 39(3). DOI:10.1111/cgf.14034.
26. Adadi A and Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018; 6: 52138–52160. DOI:10.1109/ACCESS.2018.2870052.
27. Zhao X, Wu Y, Lee DL et al. IForest: Interpreting Random Forests via Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 2019; DOI:10.1109/TVCG.2018.2864475.
28. Kahng M, Thorat N, Chau DHP et al. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics* 2019; DOI:10.1109/TVCG.2018.2864500. 1809.01587.
29. Wang J, Gou L, Shen HW et al. DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks. *IEEE Trans Vis Comput Graph* 2019; DOI:10.1109/TVCG.2018.2864504.
30. Kahng M, Andrews PY, Kalro A et al. ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models. *IEEE Transactions on Visualization and Computer Graphics* 2018; 24(1): 88–97. DOI:10.1109/TVCG.2017.2744718. 1704.01942.
31. Liu M, Liu S, Su H et al. Analyzing the Noise Robustness of Deep Neural Networks. In *Proceedings of the 2018 IEEE Conference on Visual Analytics Science and Technology*. ISBN 9781538668610. DOI:10.1109/VAST.2018.8802509. 2001.09395.
32. Ming Y, Cao S, Zhang R et al. Understanding Hidden Memories of Recurrent Neural Networks. In *Proceedings of the 2017 IEEE Conference on Visual Analytics Science and Technology*. ISBN 9781538631638. DOI:10.1109/VAST.2017.8585721. 1710.10777.
33. Spinner T, Schlegel U, Schäfer H et al. ExplAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics* 2020; DOI:10.1109/TVCG.2019.2934629. 1908.00087.
34. Ribeiro MT, Singh S and Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. ISBN 9781450342322. DOI:10.1145/2939672.2939778.
35. Lundberg SM and Lee SI. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*. DOI:10.48550/arXiv.1705.07874. 1705.07874.
36. Ming Y, Qu H and Bertini E. RuleMatrix: Visualizing and Understanding Classifiers with Rules. *IEEE Transactions on Visualization and Computer Graphics* 2019; DOI:10.1109/TVCG.2018.2864812. 1807.06228.
37. Zhang J, Wang Y, Molino P et al. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 2019; DOI:10.1109/TVCG.2018.2864499. 1808.00196.
38. Kwon BC, Eysenbach B, Verma J et al. Clustervision: Visual Supervision of Unsupervised Clustering. *IEEE Transactions on Visualization and Computer Graphics* 2018; DOI:10.1109/TVCG.2017.2745085.
39. Jaccard J, Wan CK and Turrisi R. The Detection and Interpretation of Interaction Effects Between Continuous Variables in Multiple Regression. *Multivariate Behavioral Research* 1990; DOI:10.1207/s15327906mbr2504_4.
40. Coulton C and Chow J. Interaction effects in multiple regression. *Journal of Social Service Research* 1993; DOI:10.1300/J079v16n01_09.
41. Jaccard J. Interaction effects in logistic regression. In *Research Methods*. SAGE Publications. ISBN 0761922075, 2001. DOI:10.4135/9781412984515.
42. Van Der Weele TJ and Knol MJ. A tutorial on interaction. *Epidemiologic Methods* 2014; DOI:10.1515/em-2013-0005.
43. Friedman JH and Popescu BE. Predictive learning via rule ensembles. *Annals of Applied Statistics* 2008; DOI:10.1214/07-AOAS148.
44. Greenwell BM, Boehmke BC and McCarthy AJ. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:180504755* 2018; DOI:10.48550/arXiv.1805.04755. 1805.04755.
45. Britton M. VINE: Visualizing Statistical Interactions in Black Box Models. *arXiv preprint arXiv:190400561* 2019; DOI:10.48550/arXiv.1904.00561. 1904.00561.
46. Goldstein A, Kapelner A, Bleich J et al. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics* 2015; DOI:10.1080/10618600.2014.907095. 1309.6392.
47. Lamina C, Sturm G, Kollerits B et al. Visualizing interaction effects: a proposal for presentation and interpretation. *J Clinical Epidemiology* 2012; 65(8): 855–862.
48. Hooker G. Discovering additive structure in black box functions. In *Proc. ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining*. pp. 575–580. DOI:10.1145/1014052.1014122.
49. Amer M, Daim TU and Jetter A. A review of scenario planning. *Futures* 2013; DOI:10.1016/j.futures.2012.10.003.
50. Harries C. Correspondence to what? Coherence to what? What is good scenario-based decision making? *Technological Forecasting and Social Change* 2003; DOI:10.1016/S0040-1625(03)00023-4.
51. Golfarelli M, Rizzi S and Proli A. Designing what-if analysis: Towards a methodology. In *DOLAP: Proc. ACM International Workshop on Data Warehousing and OLAP*. ISBN 1595935304. DOI:10.1145/1183512.1183523.
52. Barnes CD, Quiason JL, Benson C et al. Success stories in simulation in health care. In *Winter Simulation Conference Proceedings*. DOI:10.1145/268437.268772.
53. Bohanec M, Zupan B and Rajkovič V. Applications of qualitative multi-attribute decision models in health care. *Int J Medical Informatics* 2000; 58-59. DOI:10.1016/S1386-5056(00)00087-3.

54. Card AJ, Ward JR and Clarkson PJ. Beyond FMEA: The structured what-if technique (SWIFT). *Journal of Healthcare Risk Management* 2012; 31(4): 23–29. DOI:https://doi.org/10.1002/jhrm.20101. https://onlinelibrary.wiley.com/doi/pdf/10.1002/jhrm.20101.
55. Swart RJ, Raskin P and Robinson J. The problem of the future: Sustainability science and scenario analysis. *Global Environmental Change* 2004; 14(2). DOI:10.1016/j.gloenvcha.2003.10.002.
56. Reilly M and Willenbockel D. Managing uncertainty: A review of food system scenario analysis and modelling. *Philosophical Trans Royal Soc B: Biological Sciences* 2010; 365(1554). DOI:10.1098/rstb.2010.0141.
57. van Sluisveld MA, Hof AF, Carrara S et al. Aligning integrated assessment modelling with socio-technical transition insights: An application to low-carbon energy scenario analysis in Europe. *Technological Forecasting and Social Change* 2020; 151. DOI:10.1016/j.techfore.2017.10.024.
58. Salas E, Priest HA, Wilson KA et al. Scenario-Based Training: Improving Military Mission Performance and Adaptability. In *Military life: The psychology of serving in peace and combat (Vol. 2): Operational Stress*. Praeger Security International, 2006. pp. 32–53.
59. Karvetski CW, Lambert JH and Linkovz I. Scenario and multiple criteria decision analysis for energy and environmental security of military and industrial installations. *Integrated Environmental Assessment and Management* 2011; 7(2). DOI:10.1002/ieam.137.
60. Huss WR and Honton EJ. Scenario planning-What style should you use? *Long Range Planning* 1987; 20(4). DOI:10.1016/0024-6301(87)90152-X.
61. Huss WR. A move toward scenario analysis. *International Journal of Forecasting* 1988; 4(3). DOI:10.1016/0169-2070(88)90105-7.
62. Krause J, Perer A and Ng K. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. pp. 5686–5697. DOI:10.1145/2858036.2858529.
63. Wexler J, Pushkarna M, Bolukbasi T et al. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics* 2020; DOI:10.1109/TVCG.2019.2934619. 1907.04135.
64. Sohns JT, Garth C and Leitte H. Decision boundary visualization for counterfactual reasoning. In *Computer Graphics Forum*. Wiley Online Library. DOI:10.1111/cgf.14650.
65. Molnar C. *Interpretable Machine Learning*. 2 ed. 2022. URL https://christophm.github.io/interpretable-ml-book.
66. Edwards W. The theory of decision making. *Psychological bulletin* 1954; 51(4): 380. DOI:10.1037/h0053870.
67. Slovic P, Lichtenstein S and Fischhoff B. Decision making. In *Stevens' handbook of experimental psychology 2nd ed*, volume 2. Wiley, 1988. pp. 673–738. URL http://hdl.handle.net/1794/22321.
68. Kahneman D and Tversky A. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 2013. pp. 99–127. DOI:10.1142/9789814417358_0006.
69. Tversky A and Kahneman D. The framing of decisions and the psychology of choice. In *Behavioral decision making*. Springer, 1985. pp. 25–41. DOI:10.1126/science.7455683.
70. Tversky A and Kahneman D. Rational choice and the framing of decisions. In *Multiple criteria decision making and risk analysis using microcomputers*. Springer, 1989. pp. 81–126. DOI:https://doi.org/10.1007/978-3-642-74919-3_4.
71. Kahneman D, Slovic SP, Slovic P et al. *Judgment under uncertainty: Heuristics and biases*. Cambridge university press, 1982. ISBN 0 521 284147.
72. Milkman KL, Chugh D and Bazerman MH. How can decision making be improved? *Perspectives on psychological science* 2009; 4(4): 379–383. DOI:10.1111/j.1745-6924.2009.01142.x.
73. Wilson HJ and Daugherty PR. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review* 2018; 96(4): 114–123.
74. Bastani H, Bastani O and Sinchaisri WP. Improving human decision-making with machine learning. *arXiv preprint arXiv:210808454* 2021; DOI:10.48550/arXiv.2108.08454.
75. Khosravi K, Shahabi H, Pham BT et al. A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods. *Journal of Hydrology* 2019; 573: 311–323. DOI:10.1016/j.jhydrol.2019.03.073.
76. Secinaro S, Calandra D, Secinaro A et al. The role of artificial intelligence in healthcare: a structured literature review. *BMC Medical Informatics and Decision Making* 2021; 21(1): 1–23. DOI:10.1186/s12911-021-01488-9.
77. Veropoulos K. *Machine learning approaches to medical decision making*. PhD Thesis, University of Bristol, England, 2001.
78. Sahoo AK, Pradhan C and Das H. Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making. In *Nature Inspired Computing for Data Science*. Springer, 2020. pp. 201–212. DOI:10.1007/978-3-030-33820-6_8.
79. Thirumalai C, Duba A and Reddy R. Decision making system using machine learning and pearson for heart attack. In *International Conference on Electronics, Communication and Aerospace Technology*, volume 2. IEEE, pp. 206–210. DOI:10.1109/ICECA.2017.8212797.
80. Duan Y, Edwards JS and Dwivedi YK. Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. *International Journal of Information Management* 2019; 48: 63–71. DOI:10.1016/j.ijinfomgt.2019.01.021.
81. Dwivedi YK, Hughes L, Ismagilova E et al. Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int J Information Management* 2021; 57: 101994. DOI:10.1016/j.ijinfomgt.2019.08.002.
82. Simon S, Mittelstädt S, Keim DA et al. Bridging the Gap of Domain and Visualization Experts with a Liaison. In *Eurographics Conference on Visualization (EuroVis) - Short Papers*. The Eurographics Association. DOI:10.2312/eurovisshort.20151137.
83. Munzner T. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics* 2009; DOI:10.1109/TVCG.2009.111.

84. Aboutorabi H and Kung L. Application of coupled structural acoustic analysis and sensitivity calculations to a tire noise problem. *Tire Science and Technology* 2012; 40(1): 25–41. DOI:https://doi.org/10.2346/1.3684489.
85. Lu Y, Garcia R, Hansen B et al. The State-of-the-Art in Predictive Visual Analytics. *Computer Graphics Forum* 2017; 36(3). DOI:10.1111/cgf.13210.
86. Sacha D, Kraus M, Keim DA et al. VIS4ML: An Ontology for Visual Analytics Assisted Machine Learning. *IEEE Trans Vis Comput Graph* 2019; DOI:10.1109/TVCG.2018.2864838.
87. Zhou J, Arshad SZ, Luo S et al. Effects of uncertainty and cognitive load on user trust in predictive decision making. In *Human-Computer Interaction–INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part IV 16*. Springer, pp. 23–39.
88. Alin A. Minitab. *Wiley Interdisciplinary Reviews: Computational Statistics* 2010; DOI:10.1002/wics.113.
89. *JMP 12 Basic Analysis*, 2015.
90. Newman GE and Scholl BJ. Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin and Review* 2012; 19(4). DOI:10.3758/s13423-012-0247-5.
91. Coppers S, Luyten K, Vanacken D et al. Fortunettes: Feedforward about the future state of gui widgets. *Proc ACM Hum-Comput Interact* 2019; 3(EICS). DOI:10.1145/3331162.
92. Gutiérrez-Gómez L, Petry F and Khadraoui D. A comparison framework of machine learning algorithms for mixed-type variables datasets: a case study on tire-performances prediction. *IEEE Access* 2020; 8: 214902–214914. DOI:10.1109/ACCESS.2020.3041367.
93. Gleicher M, Albers D, Walker R et al. Visual comparison for information visualization. *Information Visualization* 2011; 10(4): 289–309. DOI:10.1177/1473871611416549.
94. Piñeiro G, Perelman S, Guerschman JP et al. How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecological Modelling* 2008; 216(3-4). DOI:10.1016/j.ecolmodel.2008.05.006.
95. Ciesielski K. *Set Theory for the Working Mathematician*. London Mathematical Soc. Student Texts, Cambridge University Press, 1997. DOI:10.1017/CBO9781139173131.
96. du Prel JB, Hommel G, Röhrig B et al. Confidence interval or p-value? *Deutsches Arzteblatt international* 2009; DOI:10.3238\%2Farztebl.2009.0335.
97. Harter HL. The Method of Least Squares and Some Alternatives: Part V. *International Statistical Review* 1975; DOI:10.2307/1403110.
98. Fox J and Weisberg S. Fitting Linear Models. *JMP, A Business Unit of SAS* 2011; .
99. Ghoniem M, Fekete JD and Castagliola P. A comparison of the readability of graphs using node-link and matrix-based representations. In *Proceedings of the IEEE Symposium on Information Visusalization, INFOVIS*. DOI:10.1109/INFVIS.2004.1.
100. Boos DD and Stefanski LA. P-value precision and reproducibility. *American Statistician* 2011; DOI:10.1198/tas.2011.10129.
101. Huber PJ. Projection Pursuit. *The Annals of Statistics* 2007; 13(2). DOI:10.1214/aos/1176349519.
102. Pearson K. Notes on the History of Correlation. *Biometrika* 1920; DOI:10.2307/2331722.
103. Siramdasu Y, Li K and Wheeler R. Understanding Tire Dynamic Characteristics for Vehicle Dynamics Ride Using Simulation Methods. *Tire Science and Technology* 2020; 48(3): 188–206. DOI:10.2346/tire.19.180196.
104. Merkel D. Docker: lightweight linux containers for consistent development and deployment. *Linux journal* 2014; 2014(239): 2. URL `https://www.seltzer.com/margo/teaching/CS508.19/papers/merkel14.pdf`.
105. Lam H, Bertini E, Isenberg P et al. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics* 2012; 18(9): 1520–1536. DOI:10.1109/TVCG.2011.279.
106. Saraiya P, North C and Duca K. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Trans Vis Comput Graph* 2005; 11(4): 443–456. DOI:10.1109/TVCG.2005.53.
107. Stasko J. Value-driven evaluation of visualizations. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. BELIV '14, ACM. ISBN 9781450332095, p. 46–53. DOI:10.1145/2669557.2669579.
108. Isenberg T, Isenberg P, Chen J et al. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 2818–2827. DOI:10.1109/TVCG.2013.126.
109. Elmqvist N and Yi JS. Patterns for visualization evaluation. *Information Visualization* 2015; 14(3): 250–269. DOI:10.1177/1473871613513228.
110. Carpendale S. Evaluating information visualizations. In *Lecture Notes in Computer Science*, volume 4950. Springer, Berlin, Heidelberg. ISBN 354070955X, pp. 19–45. DOI:10.1007/978-3-540-70956-5_2.
111. Weber GH, Carpendale S, Ebert D et al. Apply or die: On the role and assessment of application papers in visualization. *IEEE Computer Graphics and Applications* 2017; 37(3): 96–104. DOI:10.1109/MCG.2017.51.
112. Mazza R and Berre A. Focus group methodology for evaluating information visualization techniques and tools. In *Proceedings of the International Conference on Information Visualization (IV'07)*. IEEE, pp. 74–80. DOI:10.1109/IV.2007.51.
113. Erlingsson C and Brysiewicz P. A hands-on guide to doing content analysis. *African Journal of Emergency Medicine* 2017; 7(3): 93–99. DOI:10.1016/j.afjem.2017.08.001.
114. Janesick VJ. Peer debriefing. *The Blackwell Encycl Sociol* 2007; DOI:10.1002/9781405165518.wbeosp014.pub2.
115. Joshi A, Kale S, Chandel S et al. Likert scale: Explored and explained. *British Journal of Applied Science and Technology* 2015; 7(4): 396. DOI:10.9734/BJAST/2015/14975.
116. Jamieson S. Likert scales: How to (ab) use them? *Medical education* 2004; 38(12): 1217–1218. DOI:10.1111/j.1365-2929.2004.02012.x.
117. Harpe SE. How to analyze likert and other rating scale data. *Currents in Pharmacy Teaching and Learning* 2015; 7(6): 836–850. DOI:10.1016/j.cptl.2015.08.001.
118. Bishop PA and Herron RL. Use and misuse of the likert item responses and other ordinal measures. *International Journal of Exercise Science* 2015; 8(3): 297. URL `https://pubmed.ncbi.nlm.nih.gov/27182418/`.

119. Ross A and Willson VL. *One-Sample T-Test*. Rotterdam: SensePublishers. ISBN 978-94-6351-086-8, 2017. pp. 9–12. DOI:10.1007/978-94-6351-086-8_2. URL https://doi.org/10.1007/978-94-6351-086-8_2.
120. Sprent P. *Sign Test*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-04898-2, 2011. pp. 1316–1317. DOI:10.1007/978-3-642-04898-2_515. URL https://doi.org/10.1007/978-3-642-04898-2_515.
121. Taylor J. *Classical Mechanics*. G - Reference,Information and Interdisciplinary Subjects Series, University Science Books, 2005. ISBN 9781891389221.
122. Nagaya K, Ikai S, Chiba M et al. Tire with self-repairing mechanism. *JSME Int J Ser C Mechanical Syst, Machine Elements and Manufacturing* 2006; 49(2): 379–384. DOI: 10.1299/jsmec.49.379.
123. Student. The probable error of a mean. *Biometrika* 1908; : 1–25DOI:https://doi.org/10.1007/978-1-4612-4380-9_4. URL https://www.york.ac.uk/depts/maths/histstat/student.pdf.
124. De Winter JC. Using the student's t-test with extremely small sample sizes. *Practical Assessment, Research, and Evaluation* 2013; 18(1): 10. DOI:10.7275/e4r6-dj05.
125. Tavazza F, DeCost B and Choudhary K. Uncertainty prediction for machine learning models of material properties. *ACS omega* 2021; 6(48): 32431–32440.
126. Guyon I et al. Practical feature selection: from correlation to causality. *Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security* 2008; : 27–43.
127. De Gonzalez AB, Cox DR et al. Interpretation of interaction: A review. *Annals of Applied Statistics* 2007; 1(2): 371–385. DOI:10.1214/07-AOAS124.
128. Chakhchoukh MR, Boukhelifa N and Bezerianos A. Understanding how in-visualization provenance can support trade-off analysis. *IEEE Transactions on Visualization and Computer Graphics* 2022; .
129. Gotz D and Zhou MX. Characterizing users' visual analytic activity for insight provenance. In *2008 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, pp. 123–130.
130. Segel E and Heer J. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics* 2010; 16(6): 1139–1148.
131. Stijven S, Minnebo W and Vladislavleva K. Separating the wheat from the chaff: on feature selection and feature importance in regression random forests and symbolic regression. In *Proceedings of the Annual Conference Companion Genetic Evolutionary Computing*. pp. 623–630. DOI:10.1145/2001858.2002059.
132. Sedlmair M, Meyer M and Munzner T. Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans Vis Comput Graph* 2012; 18(12): 2431–2440. DOI:10.1109/TVCG.2012.213.
133. Pajer S, Streit M, Torsney-Weir T et al. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE Trans on Vis Comput Graph* 2016; 23(1): 611–620. DOI:10.1109/TVCG.2016.2598589.
134. Laurillau Y, Nguyen VB, Coutaz J et al. The TOP-slider for multi-criteria decision making by non-specialists. In *ACM Int. Conf. Proceeding Series*. NordiCHI '18, ACM. ISBN 9781450364379, pp. 642–653. DOI:10.1145/3240167.3240185.
135. Nadareishvili I, Mitra R, McLarty M et al. *Microservice architecture: aligning principles, practices, and culture*. " O'Reilly Media, Inc.", 2016.
136. vanden Broucke S. Discovering Interaction Effects in Ensemble Models, 2019. URL https://blog.macuyiko.com/post/2019/discovering-interaction-effects-in-ensemble-models.html.

The supplemental material, including the demo video is located at:

https://osf.io/94bzf?view_only=5da9f4a35d0c48c9963e13fb576ec1e6

The Latex source code is located at:

https://osf.io/7tgmp/?view_only=49b7d66f5a384fa7a0ca68362aa1cab1