Visualization Guidelines for Model Performance Communication Between Data Scientists and Subject Matter Experts

Ashley Suh, Gabriel Appleby, Erik W. Anderson, Luca Finelli, Remco Chang, Dylan Cashman

Abstract— Presenting the complexities of a model's performance is a communication bottleneck that threatens collaborations between data scientists and subject matter experts. Accuracy and error metrics alone fail to tell the whole story of a model – its risks, strengths, and limitations – making it difficult for subject matter experts to feel confident in deciding to use a model. As a result, models may fail in unexpected innexpected ways if their weaknesses are not clearly understood. Alternatively, models may go unused, as subject matter experts disregard poorly presented models in favor of familiar, yet arguably substandard methods. In this paper, we propose effective use of visualization as a medium for communication between data scientists and subject matter experts. Our research addresses the gap between common practices in model performance communication and the understanding of subject matter experts and decision makers. We derive a set of communication guidelines and recommended visualizations for communicating model performance based on interviews of both data scientists and subject matter experts at the same organization. We conduct a follow-up study with subject matter experts to evaluate the efficacy of our guidelines in presentations of model performance with and without our recommendations. We find that our proposed guidelines made subject matter experts more aware of the tradeoffs of the presented model. Participants realized that current communication methods left them without a robust understanding of the model's performance, potentially giving them misplaced confidence in the use of the model.

Index Terms—Visual Communication, Regression Models, Model Performance

Based on regression, not classification

1 Introduction

SMR: Sloan Management Review

The widespread use of artificial intelligence (AI) has reached far beyond academia: in a recent study with over 70 Fortune 1000 companies, the percentage of firms investing at least $50 million in AI increased by nearly 63% from 2018 to 2020 [1]. Surprisingly though, only 15% of firms reported that they had *actually deployed* AI capabilities into widespread production. Similar findings were echoed by Ransbotham et al. in an MIT SMR report: 7 out of 10 companies reported minimal to no impact from their organization's AI use, with only 10% of companies reporting significant financial benefits from implementing AI [47].

It is undeniable that the potential benefits of AI and machine learning (ML) have driven great interest across a variety of domains such as healthcare, biology, and commercial industries [22, 31, 60]. It is also apparent that there is still a practical gap between the promise, and the actual execution [1]. Previous research has investigated how practitioners can improve AI and ML collaborations [29, 46], and poor communication is consistently cited as a top obstacle in model development between data scientists and subject matter experts [5]. While previous visualization research has contributed many methods for helping a data scientist develop a better model [49], we believe that a more fundamental question needs to be answered once a model has been developed: How can we best communicate these models to subject matter experts, who ultimately decide what to do with it?

To investigate how issues in communication can lead to downstream issues with model deployment, we conducted a series of interviews with both data science practitioners who develop models and the subject matter experts (SMEs) that must work with the outputs of those models. We present the results of these interviews to identify best practices and challenges associated with model performance communication. We then leverage these results to create a set of prescriptive guidelines designed to facilitate multi-disciplinary model communication through visualization and presentation.

Communication between members of different disciplines is a difficult task that is often facilitated by various visualization methods [7, 36, 41, 56]. Although data visualization is a common tool employed by data scientists when presenting model performance results, our interviews revealed that many commonplace predictive model visualizations (*e.g.*, residual plots) may result in greater confusion, particularly when the audience is not familiar with technical data science concepts or visualization methods. To combat these misunderstandings, data scientists often resort to presenting common performance metrics, such as explained variance or mean squared error, to illustrate their model's performance in a presentation [53]. We find that the presentation of these metrics alone is insufficient for SMEs who must make decisions and recommendations based on the model's outputs.

We analyzed our interviews both quantitatively (using iterative coding) and qualitatively (using NLP methods) to deduce the types of communication difficulties that occur when a data scientist presents a model to subject matter experts. We find that while data scientists and SMEs are often concerned about similar issues (*e.g.*, strengths and weaknesses of a model, edge cases, etc.), each person approaches these issues drawing upon very different sets of context and experiences, and even communicates using domain-specific vocabulary that may or may not be shared between groups. From the analysis of our interviews, we distill guidelines for data scientists to present the complexities of model performance to SMEs. We then validate the effectiveness of our proposed guidelines in a followup study focusing on SMEs. Our results show that current data science practices give SMEs a misplaced sense of confidence in a model's performance. However, we demonstrate that SMEs presented with model performance data using our guidelines were able to better understand the model's strengths and weaknesses, enabling them to use their subject expertise to make a more calculating interpretation of the model.

They have different backgrounds

We organize the remainder of this manuscript as follows. After an overview of related work in Section 2, we discuss the interview process and its goals in Section 3. Section 4 details the results of our interview study, which inform our prescriptive guidelines in Section 5. We then describe the validation study performed in Section 6 before providing a discussion on both the institutional impact of the guidelines and directions for future work in Section 7.

In summary, our major contributions are as follows:

- An interview study with data scientists and SMEs who regularly collaborate on regression models.
- Guidelines that emerge from the analysis of our interviews and a review of prior literature.
- A validation of our proposed visualization and communication guidelines with SMEs.

- *Ashley Suh, Gabriel Appleby, and Remco Chang are with Tufts University. E-mail: {ashley.suh, gabriel.appleby, remco.chang}@tufts.edu*
- *Erik W. Anderson, Luca Finelli, and Dylan Cashman are with Novartis Pharmaceuticals Corporation, Data Science and AI. E-mail: {erik.anderson, luca.finelli, dylan.cashman}@novartis.com.*

2 Related Work

SME or domain scientist

For the remainder of this work, we broadly refer to any practitioner who works on and builds predictive models as a *data scientist*. We refer to *subject matter experts (SMEs)* as model consumers with primary roles and expertise outside of data science and machine learning. We differentiate between model consumers who take the role as 'executives' (*i.e.* stakeholders) and those whose primary function is as a domain scientist (SME) that works with data scientists to develop and deploy predictive models in their daily work. Although some SMEs may also be stakeholders, we make the distinction in this paper that an SME has a specialized set of knowledge for a particular scientific domain.

2.1 Visualization for Model Interpretability

Effective presentation of a predictive model's performance to domain scientists, SMEs, and other stakeholders is an active topic of research in visualization. Researchers studying explainable AI seek to help users interpret and explain the inferences of AI models by visualizing the internal workings of those models [16, 40, 43, 61]. Metrics and principles are posed for explainable AI [26, 48], guidelines for defining *interpretability* are suggested [12, 63], and visual analytic tools can enhance machine learning and AI transparency [9, 28, 37, 49]. Hohman et al. conducted an interview study with machine learning experts to understand how interactive interfaces could better support model interpretation for data scientists [27]. We observe that a majority of ML interpretability support is targeted towards supporting experts (*e.g.*, [49]) and not necessarily the stakeholders of ML.

While many of these works are relevant to the research presented here, the concept of explainability is discussed at too low of a level. The proposed solutions are typically complex, often require training, and may not be semantically meaningful for the SME audience. From our interview study (Section 3), we found that solutions were needed to facilitate communication between data scientists and SMEs at a higher level. To this end, we identify what data scientists and SMEs each find most valuable in the interpretation of a model's outcome, and propose visualization guidelines (Section 5) that can be used broadly by practitioners for communicating model performance.

2.2 Data Science & ML in Practice

The desire and demand for AI/ML at organizations outside of big tech is well documented [1, 30]. Consequently, there has been a sharp increase in the use (and subsequent risk) of advanced analytics in nontraditional ML domains (*e.g.*, for healthcare and childcare) [2, 18, 25, 64]. However, when introducing AI/ML techniques in practice, it is also necessary to investigate the risks, limitations, edge cases, and weaknesses of a predictive model before its deployment [2, 18, 25]. Although strides are continuously being made to improve the transparency of ML models, empirical research shows these techniques are rarely deployed in high stakes domains [11, 17, 24, 54]. Close to our work is the recent design study ~~with~~ by Zytek et al. [64], in which the authors collaborate with SMEs that use ML techniques in child welfare. In Zytek et al.'s work, an iterative design process combines ML practitioners and childcare experts to identify key challenges in their workflow. The final result is a visual analytics tool to help alleviate interpretability challenges for the domain experts. In contrast to Zytek et al.'s work and similar characterizations of ML interpretability challenges, we instead focus on communication bottlenecks between data scientists and SMEs that prevent model assessment and deployment.

Challenges in ML collaborations directly affect the eventual production of AI/ML solutions, particularly when the model's performance can not be well understood. Seneviratne et al. argue that work is needed to bridge the implementation gap of machine learning by merging ML algorithms into the 'socio-technical' milieu of the organization [52]. Similar work describes widespread systematic issues in the adaption of predictive models at healthcare organizations, citing the mismatch between stakeholders and their understanding of technological innovations [11]. Shah et al. suggests that the utility of ML algorithms could be better demonstrated in practice if the end-users could better assess the performance of a predictive model without relying on standard performance metrics [53]. In this work, we intentionally study how the performance of a predictive model can be effectively communicated to SMEs through visualization. Visualization and visual communication is often deployed to bridge communication and interpretability gaps, particularly when an audience may not have a similarly technical background [7, 36, 56]. This reasoning has been instrumental in guiding us to improve the accessibility of predictive models through the careful application of visualization for data science and SME communication.

Milieu: a person's social environment

2.3 Characterizing ML Collaborations

Across enterprise surveys that highlight common challenges for AI/ML collaborations, communication is continuously cited as the most difficult to overcome [1]. Research in ML, visual analytics, and human factors have attempted to characterize the precise nature of challenges surrounding collaborations between machine learning researchers and domain scientists. After 6 months of fieldwork with a corporate data science team, Passi and Jackson found that data science practitioners had several frustrations concerning the communication with stakeholders [46], such as the lack of trust from stakeholders when the results or inner workings of an ML model were being discussed. These ML collaboration challenges exist in other sectors outside of data science as well. Hopkins and Booth conducted an interview study with stakeholders from organizations outside of 'Big Tech' to understand how resource constraints challenge ML workflows and development [30]. Hopkins and Booth similarly found that language barriers between ML practitioners and stakeholders prevented trust and deployment of models and data science results.

Language barriers between data scientists and SMEs can occur at many stages of the ML collaboration process. Hong et al. conducted an interview study with ML practitioners on model interpretability and found that data scientists had difficulty assessing the prior expertise and knowledge level of SMEs when delivering results, presentations, and insights [29]. Similar in spirit to our work, Mosca et al. presented a study of client-facing data scientists to identify the common communication strategies they employ to translate the (potentially ill-defined) analysis needs of SMEs [45]. In this paper, we focus primarily on how *predictive models* (rather than general data science findings) can be best communicated with and to SMEs. To this end, we interview both data scientists and SMEs to get a sense of current gaps in their practices. Further, unlike prior work that aims to characterize and offer broad solutions to common ML collaboration challenges, we distill and validate visualization strategies that reduce the burden of presenting and communicating model performance.

In recent work that surveys previous studies on ML collaborations, Suresh et al. suggest that the end-users for ML can be characterized, and thus better understood, beyond standard *expert* versus *non-expert* roles [57]. Specifically, the authors suggest that consumers of machine learning models (*e.g.*, stakeholders, SMEs) can be categorized by their expertise (formal, instrumental, or personal) and the contexts in which this expertise manifests (ML, data domain, or the general milieu). To situate our work with Suresh et al.'s framework, we interview *subject matter experts*, not necessarily *stakeholders*, with formal knowledge of a particular data domain (*e.g.*, clinical safety or biology) and personal or instrumental knowledge of machine learning. Using Suresh et al.'s work, we are able to investigate how communication can be improved between data scientists and those with specialized knowledge in a respective data domain (SMEs).

3 Interview Study

Our work focuses directly on how communication and presentations can be improved between data scientists and subject matter experts when the end-goal is to use a predictive model. Our interview study was conducted at a large organization with two participant groups: data scientists who regularly build and communicate the performance of predictive models, and subject matter experts who make decisions using these predictive models in their daily domain-specific workflow.

3.1 Goal

The goal of our interview study was to identify and investigate challenges in how data scientists communicate model performance to SMEs

7 DS + 6 SME

PID	Role	Education	Experience with data science (1-5)	Frequency working with data (1-5)	Frequency working with regression (1-5)	Domain expertise
PID1	DS	Doctorate	(3) Familiar	(5) All day	(3) 1-3x/week	-
PID2	DS	Doctorate	(5) Expert	(4) 1-3x/day	(2) 1-3x/month	-
PID3	DS	Masters	(5) Expert	(4) 1-3x/day	(3) 1-3x/week	-
PID4	DS	Doctorate	(5) Expert	(5) All day	(3) 1-3x/week	-
PID5	DS	Masters	(4) Quite familiar	(5) All day	(2) 1-3x/month	-
PID13	DS	Doctorate	(5) Expert	(3) 1-3x/week	(5) All day	-
PID14	DS	Masters	(5) Expert	(3) 1-3x/week	(2) 1-3x/month	-
PID6	SME	Masters	(4) Quite familiar	(3) 1-3x/week	(2) 1-3x/month	Pharmacovigilance
PID7	SME	Masters	(3) Familiar	(5) All day	(3) 1-3x/week	Pharmacovigilance
PID8	SME	Bachelors	(3) Familiar	(3) 1-3x/week	(2) 1-3x/month	Quality Assurance
PID9	SME	Masters	(2) Somewhat familiar	(3) 1-3x/week	(1) Never	Commercial
PID10	SME	Masters	(3) Familiar	(5) All day	(2) 1-3x/month	Finance
PID12	SME	Masters	(4) Quite familiar	(5) All day	(3) 1-3x/week	Quality Assurance

Table 1: Demographics for our interview study in Section 3 Gender: (8) Male, (5) Female; Ages: (1) 18-29, (5) 30-39, (6) 40-49, (1) 50+.

that could be alleviated through visualization. As virtual environments become increasingly commonplace, slide-based presentations via PowerPoint are a popular mechanism by which data scientists communicate results to SMEs [21]. Therefore, we also wanted to better understand how data science presentations could be more effective across different modalities (*e.g.*, single slide, in-depth slides, interactive notebooks). Our study was also concerned with addressing challenges regarding common language barriers and conflicting vocabulary used between our two participant groups. To achieve this, all participants were shown the exact same slides and questions so we could analyze the differences in responses to the same set of questions. Finally, we considered if any visualization methods were already common and/or helpful when data scientists presented on the performance of their models.

3.2 Protocol

We solicited interview participants through email recruitment, and used *regression models* as a common baseline for the knowledge required in discussing predictive modeling. The use of regression as our model of study was motivated by the 2017–2021 Kaggle Machine Learning & Data Science Surveys[1], where regression was marked as the most common ML method used by practitioners. To this end, we chose an interpretable *and* commonly used predictive model to produce visualization guidelines that could be used broadly by practitioners.

In our email exchange, potential participants were informed that the purpose of the interview was to discuss their experiences interpreting and communicating a regression model's performance. When soliciting SMEs, we specifically asked for "SMEs who do not build models themselves, but have some experience using a regression model (looking at its predictions, deciding whether to use it, etc.) at some point in their career." In contrast, when recruiting data scientists, we targeted those that "have worked with, assessed, or communicated the performance of a regression model previously." In total, we interviewed 7 data scientists and 6 subject matter experts in varying departments and data domains. Their demographics, as well as their overall level of expertise with data science and regression modeling, can be found in Table 1.

3.3 Study Design

All of our interviews were semi-structured and lasted roughly one hour. Each interview was conducted virtually on Microsoft Teams with audio only. Shortly before each interview, participants were given a copy of the consent form which contained information about the study, its design, and their rights as participants. Each participant verbally consented to the study over a recording and was given an anonymous demographics survey to complete.

Participants were shown a set of prepared slides to walk through three different scenarios of assessing and communicating the performance of a regression model, one question at a time. Regardless of the participant's organizational role (data scientist or SME), the same questions were given during the interview. The slide deck used in our interviews, as well as all interview questions asked to participants are included as supplementary material.

All three of our interview scenarios began with at least one slide detailing any information required to answer the subsequent questions. In the first scenario, participants were asked what they would need to know about a new, "in-development" regression model before recommending its use in their workplace. During this scenario we also asked participants their preference for presentation style (single slide, in-depth, or interactive), in addition to the types of visuals and information they typically see in data science presentations. In the second scenario, participants were asked to discuss a real-world regression model they built, assessed, or communicated previously. In the last scenario, the interviewer elicited recommendations for how communication barriers could be alleviated during data science and SME collaborations.

obtained

3.4 Analysis Methodology

3.4.1 Coding

Qualitative methodology

Our coding process followed a top down approach starting with theory-driven codes as described by Braun and Clarke [8]. The process of developing our codebook was guided by literature in qualitative analysis practices [13, 42] and visualization studies [3, 33]. Our theory-based codes came from prior work studying ML workflows (see Section 2). In particular, we started with a set of codes derived from work by Suresh et al. [57], which includes objectives and tasks distilled from 58 papers across computer science and social sciences on understanding the needs of stakeholders of machine learning.

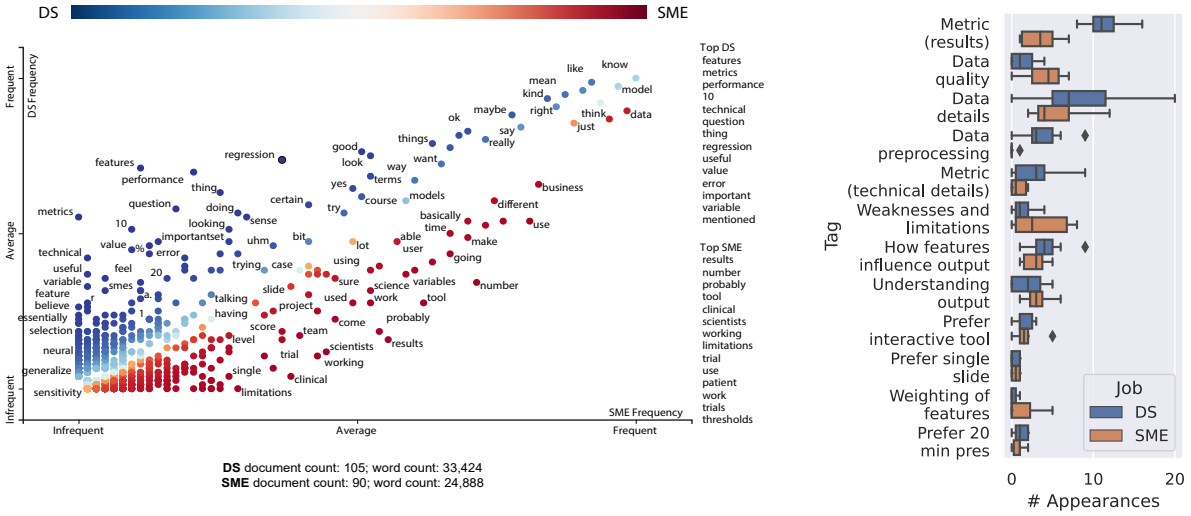
The first draft of our codebook was developed over a 2 hour working session with all authors. In addition to our theory-based codes, we added codes we considered missing based on our initial observations of the interview data. The remainder of our coding process largely followed the procedure described by Alspaugh et al. [3]. The final iteration of our codebook totaled 48 codes.

In analyzing our interviews, we drew themes from the codes that were (1) mentioned on average the most by both groups, (2) where codes frequently overlapped, and (3) where they most differed. We also considered the entropy of the codes across participants to find codes that were generally used by most participants, rather than being used often but by only one participant. The goal of this process was to perform an unbiased analysis to identify challenges, needs, and opportunities most important to both sides.

3.4.2 Scattertext Computation

In addition to a traditional analysis of code tags between classes, we also performed a discriminitive analysis between classes using Scattertext [35]. Scattertext is a tool that specializes in visualizing linguistic variation between document categories to investigate the differences between the language of the two groups. The Scattertext plot

[1] https://www.kaggle.com/c/kaggle-survey-2021



(a) Differences in word frequency between data scientists and SMEs.

(b) Frequency distribution for 12/48 codes.

Fig. 1: A summary of our computational analysis from the interview study described in Section 3. On the left, results from using Scattertext [35] on our interview documents. The x-axis represents the ranked word frequency spoken by SMEs (points shown in red), the y-axis represents ranked word frequency spoken by data scientists (points shown in blue), and words shown on the diagonal (colored closer to white) are related to both DS and SMEs. The "Top DS" and "Top SME" columns show the terms most related to each respective participant group. See Section 3.4.2 for more details. On the right, a summary of the distribution for 12 out of 48 codes derived from our interview study, sorted by their difference in means. We provide additional context and qualitative analysis on these findings in Section 4. A full summary plot of all of our codes is available in the supplemental material.

of the processed interview data can be found in Figure 1a. All text from the interviewer, stop words, and punctuation were removed.

Scattertext works for binary classes, treating the x-axis as the ranked frequency of a word within one class, and the y-axis as the ranked frequency of the word in the second class. The diagonal ($x = y$) area of the chart indicates words that SMEs and data scientists used with similar frequencies. These tend to either be words that are more connected to the interviewing environment, such as "know," and "right," or words that are inherent to the collaboration, such as "models," "data," and "explain." Areas further along the x-axis but at the bottom of the y-axis contain words used more often by SMEs (*e.g.*, "business", "limitations"). The opposite area (further along the y-axis but towards the beginning of the x-axis) demonstrate words used more often by data scientists (*e.g.*, "metrics", "features").

Each term is given a score based on its relative frequency between the two groups. The 'top' terms of each class are the highest 10 scoring terms most related to data scientists, and the lowest 10 scoring terms are most related to SMEs. The same score is then used to color each point (representing a word) using a diverging color scale from blue (more related to data scientists), white (related to both), to red (more related to SMEs). The goal of this analysis was to determine if the two groups were emphasizing different aspects of model communication, or using different terminology to describe the same concepts.

4 Findings

In this section, we present our collective findings using a thematic analysis of the differences and similarities for code usage, in addition to our findings from Scattertext (shown in Figure 1a). Our results are split into two high-level themes: (1) observed best practices for communication, and (2) identified sources of friction. We provide additional context for these emergent themes with direct quotations from interview participants. When comparing code usage, we supply the median number of times data scientists and SMEs were tagged, using M_{ds} and M_{sme} respectively. A summary of the frequencies for the 12 codes relevant to the analysis in this section can be seen in Figure 1b, a similar plot containing the frequency distribution for all 48 codes is available in the supplemental material.

4.1 What goes well: Best practices

Sharing details about the data prevents downstream issues. Many SMEs told us they provide data scientists with write-ups on unfamiliar data attributes, acronyms, or vocabulary that come directly from the SME's data domain. In general, data scientists expressed greater interest than SMEs in the technical details (*e.g.*, distributions) for the data ($M_{ds} = 7, M_{sme} = 4$) and how it was preprocessed ($M_{ds} = 3, M_{sme} = 0$). We suggest that any transformations, splits, and imputation strategies should be communicated to SMEs, as SMEs have the domain knowledge to know whether these preprocessing steps make sense. SMEs were also highly concerned about the quality of the data ($M_{ds} = 1$, $M_{sme} = 4.5$). PID06 stated that data quality issues may at first glance look like legitimate outliers from a model: "*A high number of outliers show up incorrectly . . . and when you start investigating you identify that it's because of data related or data quality issues.*"

Desire for more in-depth and interactive presentations. When asked about their preferences for presentation style, both data scientists and SMEs told us they would prefer an interactive ($M_{ds} = 1$, $M_{sme} = 1.5$) or in-depth presentation ($M_{ds} = 1$, $M_{sme} = 1$) to a single summary slide ($M_{ds} = 0$, $M_{sme} = 0.5$). 5 out of 6 SMEs stated they would prefer an interactive presentation over a short summary presentation on a model's performance, citing more comfort with the model: "*[Interactive] might get a little bit more buy in . . . if I'm indeed the end user. I think I probably would have more trust in something I've had a chance to play with.*" Data scientists initially stipulated little value in an interactive system or presentation of a model's performance, with the most common reason being that they could instead "play with the model" themselves. However, in 5 out of 6 cases, data scientists changed their mind that an interactive system would in fact help them in deciding whether to use a regression model. They particularly found this option attractive if their job was related to consulting.

Debugging models collaboratively can increase model usage. Many SMEs and data scientists noted that their collaborations are most beneficial when they can have one-on-one working sessions to debug a model and its data together. It was described to us that the data scientist would work on the model "live", present its outputs, and the SME would make

suggestions for changing the input, *e.g.*, "what if we consider doing this instead?" SMEs also noted that they would benefit from additional collaboration during a model's deployment, in addition to its lifetime after deployment, to help sustain its use in their workplace.

4.2 What goes wrong: Sources of friction

Overly technical presentations. In general, the SMEs we interviewed expressed that they had felt confused or overwhelmed by data science presentations during previous collaborations. This was largely due to unfamiliar language, metrics, and visualizations used by data scientists. In general, data scientists were much more likely to desire the technical details behind metric use than SMEs ($M_{ds} = 3$, $M_{sme} = .5$). The general metric performance code had the greatest difference of median tags across any code, with the median data scientist mentioning them 11 times, while the median SME mentioned them only 3.5 times. ScatterText (Figure 1a) reveals that data scientists were more likely to use words like "features," "regression," "metrics," and "technical." On the other hand, SMEs used more general terms such as "results," and "limitations." This mismatch in language can be intimidating during a presentation, and further the barrier between the two groups. One SME noted to us that she did not always feel comfortable asking questions during a data science presentation: "*Data scientists show a regression curve and it's so normal for them... they don't always realize that people don't understand some of the visuals for the models and what they really mean. Sometimes it just goes over your head, and I think the end-user a good chunk of the time would be too embarrassed to say - I don't get what you're talking about*"

Unfamiliar vocabulary and mismatched language. We use our results from Scattertext (Figure 1a) to guide an exploration of the differences in language and vocabulary used by participants in the interviews. Examples of mismatched language SMEs used distinctly different terminology from data scientists to describe the same concepts (*e.g.*, data scientists use "features" more often, while SMEs use "variables"). SMEs tended to use "results" holistically to describe the performance, limitations, and weaknesses of the model, while data scientists used much more specific terminology to describe model performance. SMEs used "value" in the traditional sense of the word (*e.g.*, "the model's value in my workplace"), while data scientists used "value" as a prediction or numerical value – this can lead to confusion across the two groups. Data scientists often used "error" to identify which error SMEs care about, whereas SMEs did not necessarily translate error to the overall goodness of the model.

Lack of illustrative examples for a model's limitations. SMEs were far more concerned about seeing illustrative use cases and real-life examples of the model's use. From ScatterText (Figure 1a), SMEs used terms that related to their domain application: "clinical," "trial," "patient," "user," and "business." The frequency of these words illustrates that SMEs tend to ground the discussion of a model within their domain. For building trust in a model's use, SMEs were overall interested in understanding the meaning of the output of the model ($M_{ds} = 2$, $M_{sme} = 3$), how the weighting of features affected the model ($M_{ds} = 0$, $M_{sme} = 3$), and the weaknesses and limitations of the model ($M_{ds} = 1$, $M_{sme} = 3$). When discussing what information they would like to see in a presentation, SMEs were particularly concerned about understanding the limitations of the model: "*You know, there is a human being at the end of it [a model's use]. That's the reason why I'd want as much information about the model as possible. So if I'm going to make a decision about applying this in the real world, then at least I know exactly what its limitations are before making a decision on something.*"

Frustrating **Mismatched criteria for model acceptance.** A common pain point brought up was the lack of clarity in what made a model "good" or "acceptable" to SMEs. On one hand, data scientists felt they had satisfied most of (if not all) the criteria that they established with SMEs at the start of their collaboration: *"[SMEs] don't necessarily understand the kind of undertaking you performed [improving a model]. Perhaps for you it was an awesome result of a multi month effort and you never thought you could do it. And then they [SMEs] say, isn't it still too much [error]?"* On the other hand, some SMEs expressed frustration and confusion for why a model was not performing up to their standards when all other criteria seemed to be met: *"We constantly need to revisit these kind of discussions to understand the model better ... we've given them [data scientists] everything they need - variables, all the data points. So why is the model not giving us the [expected] results?"*

5 Guidelines for Communicating Model Performance

Based on the analysis of our interviews, we derive a set of guidelines that should be followed by data scientists when presenting a model's performance to subject matter experts.

The first three guidelines relate to the choice of **visualization style** when communicating and presenting model performance to SMEs: "*Some people don't have experience with visualization outside of BBC infographics*[2]*. I do realize it can be hard for me to remove my data scientist hat and put myself into the role of somebody who's not looking at a log plot every day*" (data scientist). Guide the audience through the most important conclusions on a chart

G1: **Provide context for performance by annotating plots with stories.** Each annotation serves to decode the intended message of the visualization, beyond the visualized data and provided legend. By guiding the audience through sensible conclusions on a provided visualization, an SME could more quickly arrive at new conclusions with the same visualization, improving visualization literacy [6] over time. Use [59] for guidance on audience-based AI/ML annotations, and [51] for constructing narratives.

G2: **For any chart that communicates a model's performance, provide a range of comparisons.** SMEs found that assessing the results of a predictive model's performance is easier if it is compared against their current practices, in addition to an interpretable naive baseline model. If possible, an oracle or perfect model can also be used for comparison (or otherwise ground truth).

G3: **Visually explain the significance of aggregated metrics.** Global metrics such as explained variance or mean absolute error can seem abstract and removed from the use case. Showing metrics in visual context can help ground them; for example, visualizing the enveloping ellipse in a correlation scatterplot can give a proxy for the correlation between predicted and actual values.

The second three guidelines address concerns by both data scientists and SMEs in understanding the **caveats, edge cases, outliers, and limitations** of the model: "*If data scientists said, 'when you run these models, here is the area where we think you're going to have the most problems, or the most risk. And here's the explanation for why we think that's happening.' ... I think upfront and transparent communication about why we should expect those issues is a very big way for us to build trust and confidence in the model*" (SME).

G4: **Point to outliers in the model's performance and data space with known or plausible explanations.** The source of outliers and anomalies is often dependent on the scenario, therefore, data scientists should point SMEs to known or potential outliers, and include at least reasonable speculations behind their anomalous behavior. Use visualizations that illustrate the biggest errors in model performance, and show the distribution of features to point to potential outliers.

G5: **Describe the data used for training and testing a model, and provide examples.** The distribution, weighting, correlation, and availability of the data used in the modeling process were notable concerns from both SMEs and data scientists. Many data scientists agreed that SMEs provide essential context for the data domain, ultimately leading to improvements in model performance and transparent communication.

G6: **Communicate the limitations, weaknesses, and blind spots of the model when explaining why it should replace the existing solution(s).** SMEs want transparent information regarding the limitations of a model with both qualitative and quantitative assessments of those errors or weaknesses. Both SMEs and data scientists noted that they were able to help each other improve a model's performance once

[2] https://www.bbc.com/future/tags/infographic

the limitation of the model was fully understood, ultimately leading to higher model acceptance.

The last three guidelines address a need for **context and comfort** identified by SMEs: "*You have to make the end-user feel comfortable both in the data scientist's language, and also that if they don't understand something they can easily ask, what is this?*" (SME).

G7: **When articulating results, start slow and offer to speed up using visualization as a medium.** SMEs suggested that data scientists could spend more time highlighting aspects of their presentation that could be considered "obvious", in order to establish a common baseline for the language spoken and understood. Using visualisation to drive the discussion makes it easier to catch common misunderstandings and ground the discussion.

G8: **Provide a pre-read document with suitable explanations and illustrations of any models, metrics, or unfamiliar vocabulary.** SMEs want the option to go through model information at their own pace. A pre-read document ensures the audience can take the appropriate amount of time they need to familiarize themselves with vocabulary, visualizations, and findings before a data science presentation. Frisch et al. present details for successful pre-reads [21].

G9: **Tie in use cases for the model by illustrating real-life, objective-driven examples.** SMEs need to understand how a model's performance relates to their end-goals before recommending its use. Illustrative use cases can include individualized examples that showcase how the output or prediction of the model is affected by particular inputs, as well as examples of the model's predictions being used in real-world applications.

6 Validation Study

To validate our proposed guidelines for communicating model performance, we conduct a followup study with SMEs. The study is modeled from related work on the role of visualization in decision-making, as surveyed by Dimara and Stasko [14]. In an effort to recreate a realistic industry scenario, we presented the performance of models and asked participants to make a decision about the models and the conclusion of the presentation. We then asked for feedback on the presentation style and included visualizations.

6.1 Goals

Our study had two goals. First, we wanted to understand which guidelines were most helpful and get feedback on the visualizations that followed our guidelines. Second, we hypothesized that providing more information on model performance would increase confidence in the decision-making process, based on previous findings that communicating uncertainty improved transparency in various decision-making tasks [4, 23, 32, 39].

6.2 Method

Approximately six months after the original interview study (Section 3), the six SMEs who participated were contacted for a followup study, with four agreeing to participate. The validation study was held virtually in a one-on-one one-hour session. Each session began with a reminder of their previous participation in the study, and then proceeded through a set of slides describing two theoretical modeling scenarios. The slides are available as supplemental material, including all visualizations and data tables seen by participants.

Scenarios: Scenario 1 and scenario 2 described different datasets with similar modeling goals. In each scenario, participants were asked to listen to a presentation of a regression model's performance, and decide if they would like to use the new model being presented, or to keep the model already in use. In the first scenario, designed as a control setting, a minimal set of information was given to communicate model performance, based on common techniques described by the data scientists in our interview study, in addition to our review of the related literature. In the second scenario, the experimental setting, the presentation of model results was augmented according to our guidelines, including visualizations.

Datasets: For our first scenario, we used the Auto MPG dataset [15] to train a regression model to predict a vehicle's miles per gallon (MPG). For the second scenario, we used the California housing dataset [34] to train a regression model to predict a house's cost. All features in the original dataset were used in training our models.

XGBoost

Models used: SMEs were told that a KNN [19] model was already in use, and were asked to decide on whether to use an XGB [10] model that had been developed by a data scientist in their organization. In both modeling scenarios, the XGB model had improved performance.

Prediction tasks: Each scenario included a task that SMEs were responsible for completing with one of the two models. In scenario one, SMEs were asked to imagine they worked for a car manufacturer who would like to predict the MPG of a vehicle by its specifications before actually building the vehicle. In the second scenario, SMEs were asked to imagine that they worked for a California real estate firm that wanted to predict the value of a property before investing in it. At the end of each scenario, SMEs chose how likely they were to either keep the existing KNN model or switch to the XGB model.

Evaluation questions: At the conclusion of both scenarios, participants were asked three evaluation questions: (**Q1**) Which model would you use for the described prediction task? (**Q2**) How confident are you in your decision? (**Q3**) How confident are you in communicating why you have this decision? All three questions were rated on a 7-point Likert Scale. SMEs could respond to Q1 with one of the following ratings: definitely KNN, probably KNN, possibly KNN, not sure, possibly XGB, probably XGB, or definitely XGB. Q1 and its response choices were modeled by Guo et al.'s work on how visualizing uncertainty and alternative predictions can affect decision-making [23]. For Q2 and Q3, SMEs rated their confidence from 1 (strongly unconfident) to 7 (strongly confident). After both scenarios were completed, participants were asked to evaluate each difference in our presentations between the first and second scenario, on a scale from 1 (strongly unhelpful) to 7 (strongly helpful). SMEs gave feedback on which visualizations were most and least helpful on the same 1-7 scale.

7-point Likert scale (definitely A, probably A, possibly A, not sure, possibly B, probably B, definitely B)

6.3 Presentations

First scenario (no guidelines): At the start of both presentations, we included a short description of the dataset, the model inputs, and the SME's prediction task. An example of the input data was provided as well as a table of the first 10 prediction results for the KNN (baseline) and XGB (new) models. Following common practices described by data scientists and SMEs in our interview study, we did not include any visualizations for the first scenario's presentation. Instead, a table of common regression model metrics (R2, MSE, MAE) were displayed to convey the performance of both models.

Second scenario (with guidelines): Before starting our second scenario, we explained to SMEs that supplementary information would be included about the next scenario. This supplementary info was designed to reflect the guidelines described in Section 5; how each addition maps to guidelines is described in table 2. In total, there were four differences (**D1**-**D4**) incorporated into the presentation of our second scenario. We briefly describe each below:

- (**D1**) An initial slide, encouraging SMEs to stop us at any time to ask clarifying questions. How to do it for a real-world scenario where the number of features is much higher? Via Feature Importance?
- (**D2**) A table that defined each feature used for training and testing.
- (**D3**) High-level descriptions of the modeling algorithms and how they differ in deriving their predictions. Definitions for R2, MSE, and MAE were supplied. These definitions included how a lower or higher value for each metric translated to relative performance.
- (**D4**) Several examples of XGB performing worse than KNN, on both instance-levels and visualized regions of data (see below).

Lastly, we used 4 different annotated visualizations to illustrate differences in performance for the two models. Every visualization for XGB included a comparison visualization to the KNN (baseline) model. For each visualization, one or more annotations were included

See last page for more images

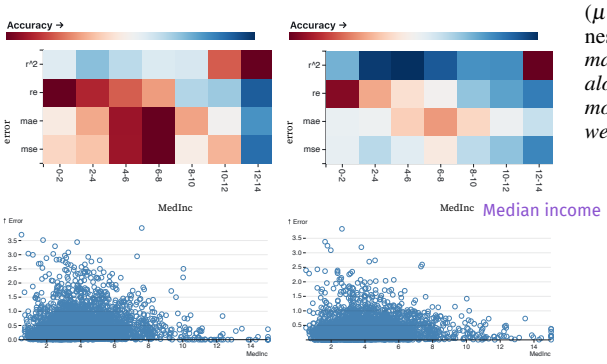
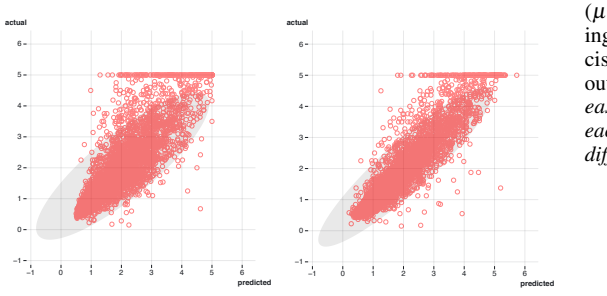
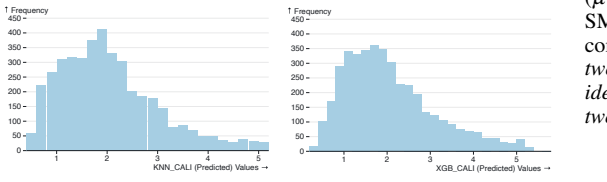
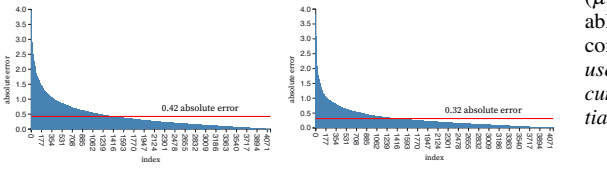
	Guidelines	How the guidelines are addressed	SME feedback (1-7) on effectiveness
D1	(G7) Ask for clarification	Slide instructing participants to ask clarifying questions about data, charts, or any language / vocabulary.	($\mu = 5.5$) Statement made the presentation feel like "*an exchange, instead of an exam.*"
D2	(G5) Data descriptions	Table detailing all data attributes used in training, with a short description and/or example for each.	($\mu = 6.5$) Quickly familiarized SMEs with the data, "*otherwise they're just numbers.*"
D3	(G8) Terminology definitions	Comparative explanation for how KNN and XGB make predictions, with definitions for R2, MSE, MAE.	($\mu = 6.25$) Gave model context: "*Without descriptions, KNN vs. XGB could be anything.*"
D4	(G6) Show model weaknesses and (G9) imply how the model would perform against real-life objectives.	Specific examples that illustrate XGB performing worse than KNN, with explanations for why.	($\mu = 6.75$) Objective-driven examples were cited as the most helpful for decision-making.
V1	(G3) Visually explain multiple metrics and (G6) see where they perform strongly or poorly. (G2) Compare performance to a baseline and (G1) see annotated examples.		($\mu = 7.0$) Easy to interpret the strengths and weaknesses of the two models: "*The visual [heatmap] made it easier to draw conclusions, the numbers alone were not as easy to translate what it [the model] means or how you could apply it if you were really using it to estimate the car's MPG.*"
V2	(G3)Visually explain R2 metric and (G4) discover outliers in prediction error. (G2) Compare performance to a baseline and (G1) see annotated examples.		($\mu = 6.5$) Useful for illuminating R2. The bounding ellipse helped discern the difference in precision for the two models, as well as identifying outliers and potential risks: "*I found it particularly easy with the ellipse in the middle to gauge where each model was better. I can see visually where the difference was, the outliers.*"
V3	(G4) Discover outliers in prediction error. (G2) Compare performance to a baseline and (G1) see annotated examples.	See the next page for a better understanding of this set of charts 	($\mu = 6.0$) Simplest visualization to grasp, helping SMEs see what types of errors the two models had compared to ground truth: "*Easy to compare the two side by side, the overall shape gives you an idea of the errors that can come up. It's just these two bars at the end each time.*"
V4	(G3)Visually explain MAE metric. (G2) Compare performance to a baseline and (G1) see annotated examples.	Bar charts (each bar is an instance) 	($\mu = 2.75$) Models were not visually distinguishable, making it unclear what finding was being communicated in the chart: "*Probably wasn't as useful in this exercise. But there are probably circumstances where it might be helpful as a differentiating figure.*"

Table 2: Table illustrating our proposed guidelines and visualizations evaluated in the 2nd scenario of our validation study with SMEs (Section 6). Each row represents a presentation element that was shown to SMEs during a PowerPoint presentation of two models' performance (KNN vs. XGB). The left column is the guidelines that are represented, the middle column describes how the guidelines are addressed, and the right column includes SME feedback for the guidelines. For the visualizations shown in the middle column, the chart on the left depicts KNN and the right depicts XGB. Participants rated the effectiveness of each presentation element on a Likert scale of 1 (very unhelpful) to 7 (very helpful).

PID	Scenario	Which model would you use?	I am confident in my decision	I am confident in communicating my decision
6	Auto MPG	Definitely XGB	7	7
	California Housing	Possibly XGB	5	4
7	Auto MPG	Possibly KNN	7	6
	California Housing	Possibly XGB	6	6
9	Auto MPG	Probably XGB	7	7
	California Housing	Probably XGB	6	7
12	Auto MPG	Probably XGB	5	5
	California Housing	Probably XGB	4	5

Table 3: Results of our validation study described in Section 6.

to point to specific differences, outliers, and conclusions between the two models. We read the annotations to SMEs during the presentation and asked if they had any follow-up questions. Examples of the 4 visualizations can be seen in Table 2, and were as follows:

- (**V1**) **Heatmap by error category**: A heatmap that illustrates aggregate errors (y-axis) across a single feature (x-axis), using a distinct color scale for each row. The heatmap allows SMEs to see how different aggregate measures differ across groups and determine if a model's performance has weaknesses along the selected feature. This is especially important when looking at how a model perform across groups of people, making sure that a model with a slight increase in overall accuracy does not sacrifice performance within a specific group. Attached to the heatmap is a one-dimensional scatterplot showing the distribution of the feature, to allow for viewing individual instances when interpreting aggregate measures.
- (**V2**) **Correlation scatterplot with confidence ellipse**: A scatterplot with predicted values from the model plotted on the x-axis, and actual values plotted on the y-axis. Behind the plotted points is a 95% confidence interval (or covariance ellipse [50]) that serves as a graphical proxy for the regression metric R2 [20]. By simultaneously displaying the aggregate metric R2 along with individual instances in the scatterplot, it provides a visual interpretation for the R2 metric. In addition, it makes it easy to see outliers in predictions and labels.
- (**V3**) **Histogram for predicted vs. actuals**: Two one-dimensional histograms, each showing the distributions of a model's prediction and actual values, respectively. By highlighting the differences in shape of predicted vs. actuals, this visualization makes it easy to see if a model might have biases (i.e. missing the high-range outliers found in the ground truth, or missing one of the modes of the distribution of actual values).
- (**V4**) **Bar chart with global error**: A bar chart where each bar represents the absolute error of a single instance; bars are sorted by descending error. This chart shows the *shape* of error by the models, and conveys how it apportions error, whether by apportioning it equally among all instances, or confining it to few. A red horizontal line is added to represent the mean absolute error, giving more context to the per instance error for the MAE and MSE metrics.

V1 and **V3** were used as part of the presentation of examples where the model performed poorly (**D4**). Any time XGB performed worse than KNN in our examples, we included a short description to explain why we believed XGB was predicting worse than KNN.

6.4 Results

Apportion: divide and allocate

First we present the feedback on our differences **D1-D4** and visualizations **V1-V4**. Then, we assess our hypothesis that providing more information about the model performance results in more confidence in decision making. Lastly, we present general qualitative feedback that came out of the experiment.

Additional context is always helpful. Participant helpfulness ratings for **D1-D4** and visualizations **V1-V4** are available in Table 2. In general, all guidelines were considered helpful, which is not surprising on its own: more information is generally always helpful in some way. Ranked by average helpfulness score, **D4** (6.75) was most helpful, reflecting our finding that understanding the strengths and weaknesses of a model, as well as seeing objective-driven examples of its predictions, impacts decision-making. It was followed by **D2** (6.5) and **D3** (6.25), which provide additional context on the raw data, the models, and the metrics. The lowest rated difference **D1** (5.5) was split, with two participants rating it as not helpful or unhelpful, and two participants rating it as strongly helpful. In sum, we found that **D1**-**D5** were all either helpful or neutral, and provide some evidence that guidelines G5, G6, G7, G8, and G9 can be partially addressed with these easy-to-implement additions.

Recount: tell someone about something

Visualized comparisons that clearly tell a story are preferred to text and numbers. Visualizations **V1** (7.0), **V2** (6.5), and **V3** (6.0) were all considered helpful, while the bar chart **V4** (2.75) was not considered helpful. **V1**, the heatmap, was strongly helpful for all participants, partly because of its easy interpretation, and partly because it was used to communicate strengths and weaknesses of the model. The scatter plot **V2** was found to be useful for understanding the metric R2, as PID12 noted that the ellipse helped them see the difference in precision for the two models, but it was also helpful for identifying outliers and alerting risks by PID6 and PID7. The histogram **V3** was considered the easiest visualization to grasp, and helped participants see what types of errors the two models had. Ratings for the bar chart **V4** ranged from neutral to strongly unhelpful, with all participants remarking that the chart did not seem visually distinguishable between the two models, and so it was not clear what the chart was trying to communicate. However, it was noted by PID7 that, for a different scenario, the chart could be helpful if the distribution of errors was very different. This suggests that visualizations are helpful in explaining model metrics, strengths and weaknesses, and outliers; however, they should only be part of a presentation if they are demonstrating a clear narrative (G1, G2, G3, G4, G6).

How to mitigate this for automatic or semi-automatic solutions?

In general, all participants provided positive feedback on the use of visualizations to communicate model performance, rather than simple text or metrics. Participants pointed out that with visualizations, they had "*quick*" or "*instant*" insights into model performance, whereas text or numbers had to be studied. Annotated examples in the visualizations (G1) were often recounted by participants as particularly helpful. However, PID9 and PID12 both noted that aside from the instant insights or annotations, they wanted to be able to pore over the visualizations outside of presentations to fully digest all of the information. They supported the usefulness of providing a pre-read document (G8), with PID9 noting: "*I'm often someone who doesn't react to something initially on a screen. . . . And when it's on the screen and I'm listening to someone present I'm back and forth between listening to the person talking and then looking at the slide. . . . I'm not digesting the visuals.*" We believe this also supports the need for more interactive tools, and we discuss this further in Section 7. We conclude that presentations should only include visualizations that tell a clear story, but that a pre-read document should be provided with a more exhaustive collection of visualizations and data, to make sure that the SME is able to ultimately decide for themselves what is relevant and what is not.

Visualizations encouraged taking a deeper look before their decision. 3 out of 4 SMEs chose the new in-development model (XGB) in the first scenario, while 4 out of 4 chose XGB in the second scenario (see Table 3). While it is encouraging that one participant switched to the more performant model (XGB), it is more interesting that all 4 SMEs reported less confidence in their decision, and 2 out of 4 reported less confidence in communicating why they had this decision in the second scenario. This does not support the original hypothesis that our participants would have increased confidence in the second scenario.

This feedback mimics previous findings in communicating uncertainty to improve transparency in decision-making [4, 32, 39]. However, it contrasts the findings from Guo et al. that visualizing uncertainty and alternative predictions improved user confidence when making decisions [23]. We believe that choosing to deploy a model may be a different type of decision than measured in previous literature, where the cost of a false negative (deploying a bad model) is very high. PID6 explained their perspective: "[In scenario two] *I can weigh each of these different parameters against everything and try to take more calculated risk. But with scenario one, I'm really not sure because I'm*

There was only one that didn't change, right? Or does this one refer to the participant who responded "Definitely XGB"?

just dependent on what limited information is provided and if it clicks, it clicks. If it doesn't, then I'm going to lose everything."

PID7 was the only participant who chose the baseline model in the first scenario, but switched to the XGB model in the second scenario based on a better understanding of the weaknesses of XGB using the visualizations and individual examples. They explained how their expertise on the model's downstream usage might dictate their choice of model: "*I might change my answer if you tell me that the problem we're trying to solve is related to old homes or really high average monthly salaries.*" Our findings suggest that the presentation of a model can make the model consumer more confident when there is a clear answer, but also less confident when there are nuances to their decision. We argue that this is the desirable effect: when there is clarity in how a model will be deployed, as PID7 explained, a more nuanced presentation can give confidence. However, when there is less clarity, it's better to be able to understand where our models perform well and perform poorly, so that risk can be calculated, rather than taking a leap of faith. This finding demonstrates the value of guidelines **G4**, **G5**, **G6**: weaving outliers, strengths, and weaknesses into the model presentation improves the SME's ability to use their own expertise in their decision-making.

7 Discussion

7.1 Institutional impact: Guidelines in practice

The guidelines presented in this work were developed in collaboration with a large pharmaceutical company. The interviews conducted as part of this work have highlighted not only that multi-disciplinary communication remains a substantial barrier to communication, but also that the effectiveness of collaboration is often over estimated. Additionally, we found that familiarity with advanced visualization techniques varies widely across users and groups, even if they are part of the same organization. This finding has highlighted the benefits gained by slowing the pace of model performance discussions by not just providing time and opportunity for SMEs to ask questions, but by encouraging them to do so.

Our industry partner has already begun to adopt these guidelines in order to address and mitigate the challenges presented in this work. The teams that are changing their reporting strategies are high-performing, multi-disciplinary groups that make regular use of machine learning models to address domain science questions. Although all collaborators have responded positively to these new guidelines, it is not possible to implement the entire suite of recommendations simultaneously. After in-depth discussions with SMEs, new visualizations will be introduced slowly to allow users the chance to become fluent in one visualization type at a time. It should be noted that introducing interactivity must be done with care. An initial attempt at providing interactive visualization techniques to the SMEs was initially met with difficulty: the SMEs needed to first adapt to new visualization methods, new interaction techniques, and new data being presented. Only after providing an interactive system that better matched their fluency with visualization techniques was the system adopted and used by the SMEs. Overall, implementation of these guidelines has resulted in a noticeable change in group dynamics: SMEs are more engaged, asking more questions, and are more positive about the collaboration.

7.2 Presentation Modalities

There are many modalities for presenting and communicating model performance. In this paper, we evaluated the use of our visualization guidelines in a PowerPoint presentation, however, SMEs told us in our interview study (Section 3) they would also find value in interactive presentations. Interestingly, when we first brought up an "interactive presentation", none of the SMEs we interviewed knew precisely what we meant. Once the interviewer explained the capabilities for an interactive system to communicate model performance, SMEs were immediately drawn to the idea that they could test various inputs to outputs themselves. One of the SMEs we interviewed mentioned that during previous model performance presentations, she has found herself wondering "*what if we tested another input instead?*" However, she felt this line of communication could be disruptive, and did not think it would be possible to see this type of interaction on-demand.

Of course, there is an indisputable trade-off cost to building interactive tools and interfaces for model performance. There is continued research being done to support effective and accessible interaction for web-based tools [55], computational notebooks [62], and visualization tools [58] that are practical for data scientists to use. Adding interaction also introduces costs for the audience (*e.g.*, false discoveries) during a presentation [38]. Further, adding interaction could be overwhelming or ineffective for an audience if they are unfamiliar with common visualization techniques [44] (see Section 7.1). To this end, we suggest introducing interaction only after SMEs show comfort and understanding of the static visualizations we have recommended.

Future work should consider the integration (and effects) of interaction into mainstream data science tools, such as pandas and sklearn. Computational notebooks seem to balance interactivity with the costs of building interactive visualization, and are historically popular with data scientists [3]. A few of the data scientists we interviewed mentioned they will design computational notebooks in which inputs to the model can be modified on-demand for SMEs during presentations, certain cells can be collapsed or left open depending on who is being presented to, and so on. As part of our contributions in this work, we designed and deployed an interactive Observable notebook [3] that has a variety of visualizations based on our interview findings.

7.3 Limitations

The requirements gathered during our interview study, in addition to the guidelines evaluated in our validation study, represent only a subset of problems and solutions from professionals collaborating within the same corporation. The scope of our studies is also narrowed to both the high funding received towards AI/ML at this corporation, and the high familiarity of data science by our interview participants. Moreover, our interviewees are highly educated professionals (4 PhD, 8 MS, 1 BS) who are considered experts in their field.

Our interview study did not attempt to quantify whether the challenges faced by our data scientists and SMEs are common across a variety of domains, skill levels, and corporations. In prior literature, studies have been done to understand AI/ML industry practices across a multitude of domains [29], in addition to sectors outside of 'Big Tech' [30]. In contrast, our study aims to characterize *how* model performance is assessed and communicated (*e.g.*, presentation styles, language used, metrics, visualizations). Our goal was to identify communication bottlenecks that prevent model usage, and suggest visualization guidelines that can alleviate these issues. While we recognize that this limits the generality of our findings, we believe that it was necessary to restrict our data scientists and SMEs to be in the same organization, in order to get both sides of the same story.

Finally, our study is focused on regression model performance communication, as opposed to general AI/ML interpretability. While it is possible that the methods suggested by our framework could be mapped to communicating any predictive model, we did not ask interview participants any questions related to other types of models. Previous work has investigated collaborative workflows on "blackbox" AI models between data scientists and stakeholders (*e.g.*, [46]) to identify common challenges. Future work should be done to assess the validity of, and extend, our guidelines to other AI/ML models as well.

8 Conclusion

We present guidelines for communicating regression model performance within a pharmaceutical organization. Based on interviews with both data scientists and subject matter experts, we identify common gaps in communication and suggest broadly applicable solutions for data scientists to use in communicating their results to SMEs. We conduct a validation study to evaluate our guidelines in practice with SMEs. Our results indicate that our guidelines were helpful in producing better understanding of the nuances of model performance, including strengths, weaknesses, and tradeoffs.

[3] https://observablehq.com/@anon1234/model-comm-vis

REFERENCES

[1] Algorithmia. 2021 enterprise trends in machine learning, 2021.

[2] J. S. Almenoff. Innovations for the future of pharmacovigilance. *Drug safety*, 30(7):631–633, 2007.

[3] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE transactions on visualization and computer graphics*, 25(1):22–31, 2018.

[4] S. Z. Arshad, J. Zhou, C. Bridon, F. Chen, and Y. Wang. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the annual meeting of the Australian special interest group for computer human interaction*, pp. 352–360, 2015.

[5] S. Berinato. Data science and the art of persuasion. *Harvard Business Review*, 97(1):126–137, 2019.

[6] K. Börner, A. Bueckle, and M. Ginda. Data visualization literacy: Definitions, conceptual frameworks, exercises, and assessments. *Proceedings of the National Academy of Sciences*, 116(6):1857–1864, 2019.

[7] M. Böttinger, H.-N. Kostis, M. Velez-Rojas, P. Rheingans, and A. Ynnerman. Reflections on visualization for broad audiences. In *Foundations of Data Visualization*, pp. 297–305. Springer, 2020.

[8] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[9] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, vol. 39, pp. 713–756. Wiley Online Library, 2020.

[10] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, New York, NY, USA, 2016. doi: 10.1145/2939672.2939785

[11] K. Cresswell and A. Sheikh. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. *International journal of medical informatics*, 82(5):e73–e86, 2013.

[12] B. Davis, M. Glenski, W. Sealy, and D. Arendt. Measure utility, gain trust: Practical advice for xai researchers. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 1–8. IEEE, 2020.

[13] J. T. DeCuir-Gunby, P. L. Marshall, and A. W. McCulloch. Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field methods*, 23(2):136–155, 2011.

[14] E. Dimara and J. Stasko. A critical reflection on visualization research: Where do decision making tasks hide? *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1128–1138, 2021.

[15] D. Donoho and E. Ramos. Primdata: data sets for use with prim-h. In *Version for second (15-18, Aug, 1983) Exposition of Statistical Graphics Technology, by American Statistical Association*, 1982.

[16] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[17] E. J. Emanuel and R. M. Wachter. Artificial intelligence in health care: will the value match the hype? *Jama*, 321(23):2281–2282, 2019.

[18] S. J. Evans. Pharmacovigilance: a science or fielding emergencies? *Statistics in medicine*, 19(23):3199–3209, 2000.

[19] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 2021/07/09/ 1989. Full publication date: Dec., 1989. doi: 10.2307/1403797

[20] M. Friendly, G. Monette, and J. Fox. Elliptical insights: understanding statistical methods through elliptical geometry. *Statistical Science*, 28(1):1–39, 2013.

[21] B. Frisch and C. Greene. What it takes to run a great virtual meeting. In *Harvard Business Review. https://hbr. org/2020/03/what-it-takes-to-run-a-great-virtual-meeting*, 2020.

[22] J. Furman and R. Seamans. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191, 2019.

[23] S. Guo, F. Du, S. Malik, E. Koh, S. Kim, Z. Liu, D. Kim, H. Zha, and N. Cao. Visualizing uncertainty and alternatives in event sequence predictions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2019.

[24] J. Harvey, G. Erdos, H. Bolam, M. A. Cox, J. N. Kennedy, and D. T. Gregory. An analysis of safety culture attitudes in a highly regulated environment. *Work & stress*, 16(1):18–36, 2002.

[25] M. Hauben and X. Zhou. Quantitative methods in pharmacovigilance. *Drug safety*, 26(3):159–186, 2003.

[26] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[27] F. Hohman, A. Head, R. Caruana, R. DeLIne, and S. Drucker. Gamut: A design probe to understand how data scientists understand machine learning models. In *SIGCHI*. ACM, May 2019.

[28] F. Hohman, A. Srinivasan, and S. M. Drucker. Telegam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*, pp. 151–155. IEEE, 2019.

[29] S. R. Hong, J. Hullman, and E. Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

[30] A. Hopkins and S. Booth. Machine learning practices outside big tech: How resource constraints challenge responsible development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 134–145, 2021.

[31] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017.

[32] S. L. Joslyn and J. E. LeClerc. Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied*, 18(1):126, 2012.

[33] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, 2012. doi: 10.1109/TVCG.2012.219

[34] R. Kelley Pace and R. Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X

[35] J. S. Kessler. Scattertext: a browser-based tool for visualizing how corpora differ. In *Proceedings of ACL-2017 System Demonstrations*. Association for Computational Linguistics, Vancouver, Canada, 2017.

[36] R. Kosara and J. Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013.

[37] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 162–172. IEEE, 2017.

[38] H. Lam. A framework of interaction costs in information visualization. *IEEE transactions on visualization and computer graphics*, 14(6):1149–1156, 2008.

[39] J. LeClerc and S. Joslyn. The cry wolf effect and weather-related decision making. *Risk analysis*, 35(3):385–395, 2015.

[40] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[41] K.-L. Ma, I. Liao, J. Frazier, H. Hauser, and H.-N. Kostis. Scientific storytelling using visualization. *IEEE Computer Graphics and Applications*, 32(1):12–19, 2011.

[42] K. M. MacQueen, E. McLellan, K. Kay, and B. Milstein. Codebook development for team-based qualitative analysis. *Cam Journal*, 10(2):31–36, 1998.

[43] S. Mohseni, N. Zarei, and E. D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*, 1, 2018.

[44] A. Mosca, A. Ottley, and R. Chang. Does interaction improve bayesian reasoning with visualization? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2021.

[45] A. Mosca, S. Robinson, M. Clarke, R. Redelmeier, S. Coates, D. Cashman, and R. Chang. Defining an Analysis: A Study of Client-Facing Data Scientists. In J. Johansson, F. Sadlo, and G. E. Marai, eds., *EuroVis 2019 - Short Papers*. The Eurographics Association, 2019. doi: 10.2312/evs.20191173

[46] S. Passi and S. J. Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–28, 2018.

[47] S. Ransbotham, S. Khodabandeh, D. Kiron, F. Candelon, M. Chu, and B. LaFountain. Expanding ai's impact with organizational learning. *MIT Sloan Management Review and Boston Consulting Group*, pp. 1–15, 2020.

[48] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand

challenges. *arXiv preprint arXiv:2103.11251*, 2021.

[49] D. Sacha, M. Kraus, D. A. Keim, and M. Chen. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE transactions on visualization and computer graphics*, 25(1):385–395, 2018.

[50] C. Schelp. An alternative way to plot the covariance ellipse. `https://carstenschelp.github.io/2018/09/14/Plot_Confidence_Ellipse_001.html`. Accessed: 2022-03-24.

[51] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics*, 16(6):1139–1148, 2010.

[52] M. G. Seneviratne, N. H. Shah, and L. Chu. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, 6(2), 2020.

[53] N. H. Shah, A. Milstein, and S. C. Bagley. Making machine learning models clinically useful. *Jama*, 322(14):1351–1352, 2019.

[54] S. Shilo, H. Rossman, and E. Segal. Axes of a revolution: challenges and promises of big data in healthcare. *Nature medicine*, 26(1):29–38, 2020.

[55] C. Sievert. *Interactive web-based data visualization with R, plotly, and shiny*. CRC Press, 2020.

[56] M. Smiciklas. *The power of infographics: Using pictures to communicate and connect with your audiences*. Que Publishing, 2012.

[57] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021. doi: 10.1145/3411764.3445088

[58] H. Suresh, K. M. Lewis, J. Guttag, and A. Satyanarayan. Intuitively assessing ml model reliability through example-based explanations and editing model inputs. In *27th International Conference on Intelligent User Interfaces*, IUI '22, p. 767–781. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3490099.3511160

[59] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–15, 2019.

[60] S. Webb. Deep learning for biology. *Nature*, 554(7690):555–558, 2018.

[61] H. J. Weerts, W. van Ipenburg, and M. Pechenizkiy. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324*, 2019.

[62] Y. Wu, J. M. Hellerstein, and A. Satyanarayan. B2: Bridging code and interactive visualization in computational notebooks. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 152–165, 2020.

[63] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020.

[64] A. Zytek, D. Liu, R. Vaithianathan, and K. Veeramachaneni. Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1161–1171, 2021.

https://observablehq.com/@anon1234/model-comm-vis