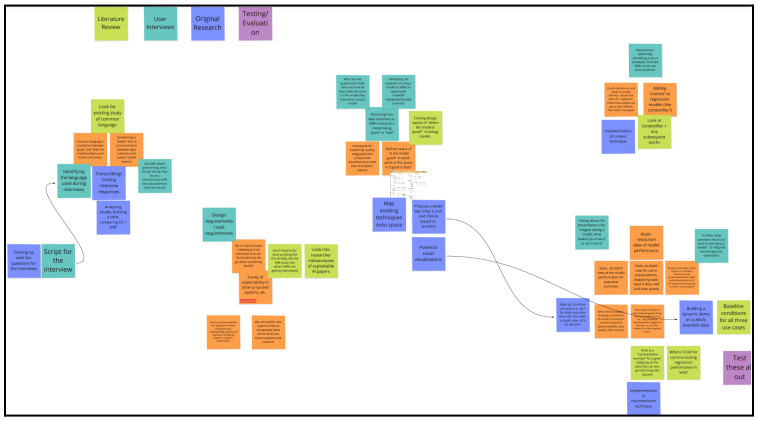
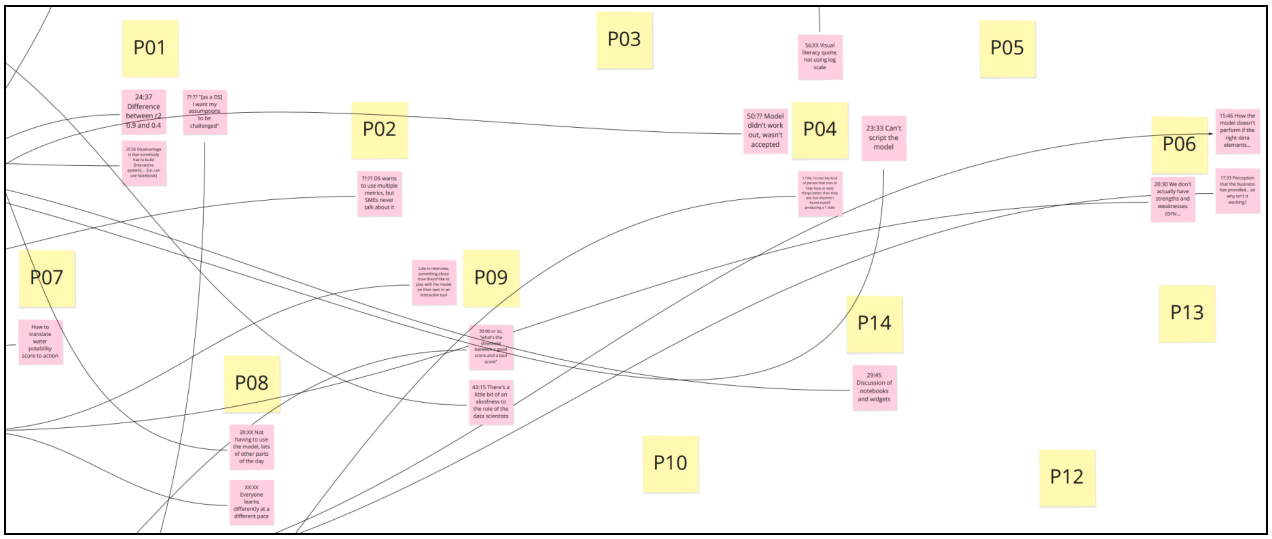
## Supplemental Material: Coding process

**Interview Analysis Process & Creation of the Guidelines: Timeline & Steps**

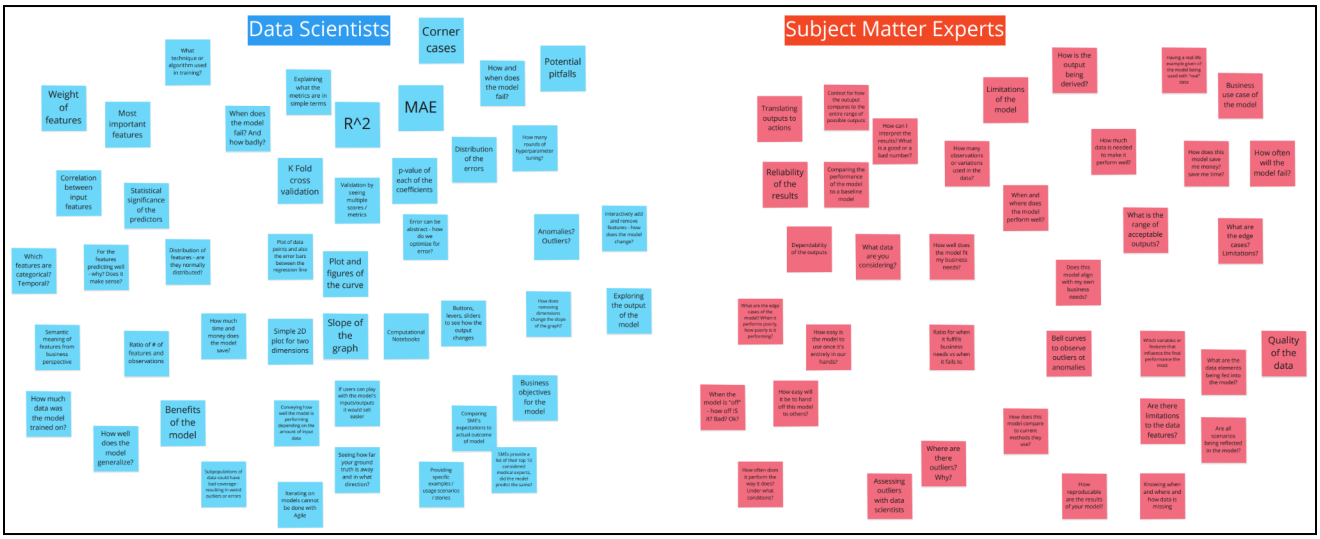
1) **Pre-study setup**: Initial brainstorming involved breaking down specific questions or concepts we wanted to cover in our interviews with data scientists & SMEs. Generally, we mapped out perceived problems in ML model communication (from prior research and field experience) to potential communication and visualization solutions.



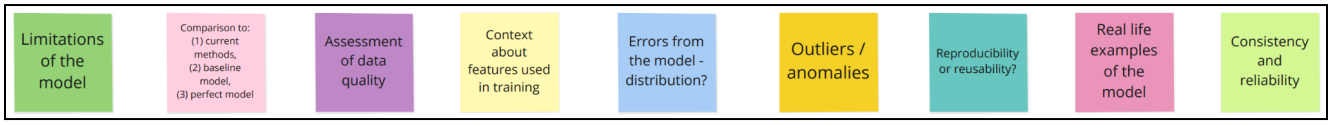

2) **Initial gathering of common, diverging, and interesting findings from interviews:** Prior to analysis, all interviews were anonymized and transcribed. While interviews were being conducted, we gathered specific timestamps of interesting quotes & notes by participants.



We also listed common mentions of model communication methods, needs, and concerns from responses of both data scientists (shown in blue) and subject matter experts (in red) – all were noted on post-it notes by the authors:

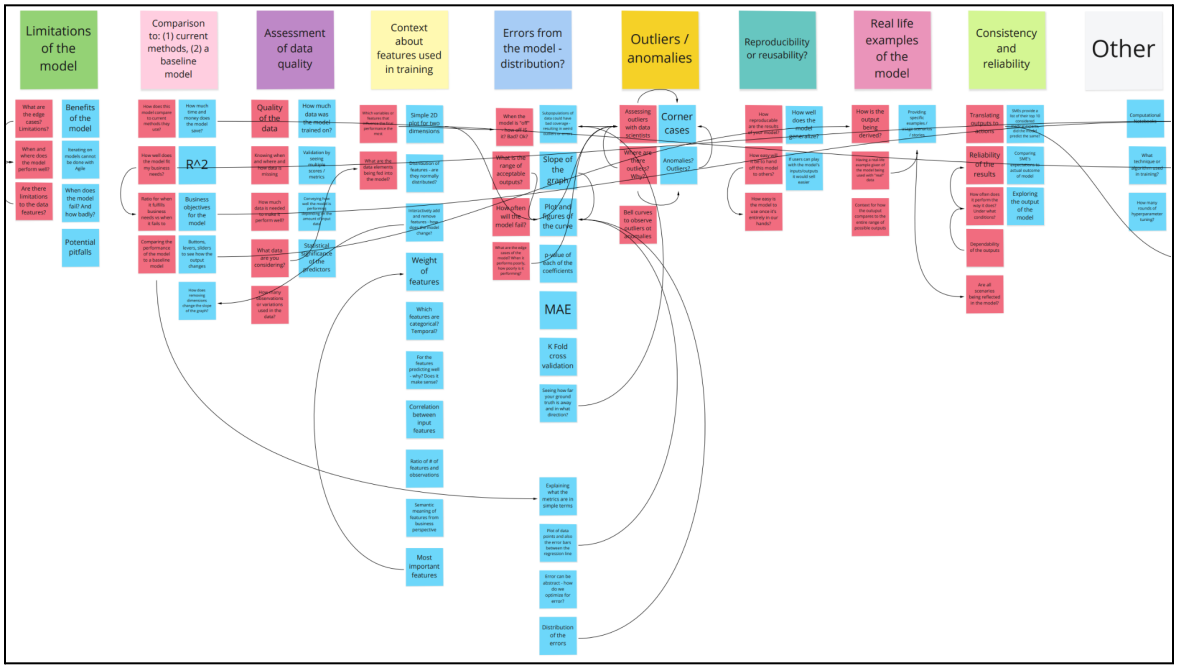


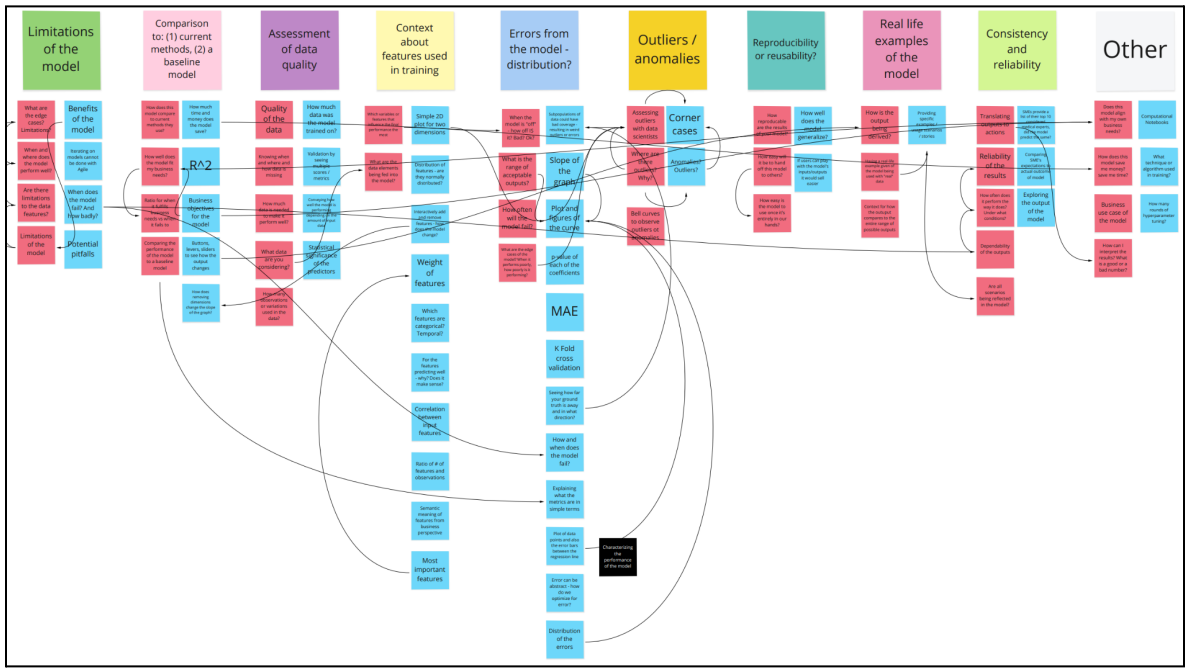

Finally, we decided on nine initial themes that we felt broadly described our collection of post-it notes, in addition to our review of prior literature in AI/ML collaborations. We consider these themes to be the first phase of our thematic analysis [*Using thematic analysis in psychology*, Virginia Braun & Victoria Clarke].




As a notable step in thematic analysis, we considered that these themes would continue to change over the progression of coding and analyzing our interviews. However, we include them here for full transparency of our thought process.

3) **Mapping concepts to themes**: Each author did a pass of mapping out our blue and red post-it notes to an appropriate theme from our collection of nine themes. We had a few reasons for doing this: first, we wanted to check whether these nine themes truly covered all concepts initially written out from data scientists and SMEs. Second, we wanted to see if data scientists or SMEs talked about a particular concept / theme more frequently. Finally, we wanted to check whether there were duplicate concepts written on post-it notes from a single group.



3) **Deciding on initial codes, developing the codebook, and coding interviews**: From this brainstorm session, the authors decided on an initial set of codes that could be used for the codebook (in addition to the theory driven codes we derived from Suresh et al. in *Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs*. Our full codebook is included in our supplemental material, including a definition of each code, inclusion and exclusion criteria when necessary, and an example quote from interview participants. We largely followed the coding process described by Alspaugh et al. in *Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices*. The timeline and procedure for our coding process is as follows:

1. All authors meet to discuss how to analyze the interviews. Authors decide on a top-down qualitative, thematic analysis approach in which interviews will be iteratively coded and categorized. Code usage frequency will be analyzed quantitatively and qualitatively.

2. All authors meet to discuss the first draft of the codebook after: (1) half of the interviews have been conducted, and (2) authors 1 and 2 have derived theory-based codes from prior literature. Participant utterances are decided as a full response made by a participant to a single interview question. If the participant changes topics mid-response, this will count as a new utterance. Multiple codes can be used per utterance.

3. Authors 1 and 2 individually do a first pass on coding the interviews over a two-week span. Both authors meet together twice to discuss any missing codes in the codebook that do not cover the interview responses. New codes are added to the codebook, inclusion and exclusion criteria are discussed in agreement with both authors.

4. All authors meet again to view the current distributions of codes, discuss missing code labels or over-used codes that should be split up. New codes are added to the codebook in agreement with all authors. Initial encapsulating themes are revised and iterated on.

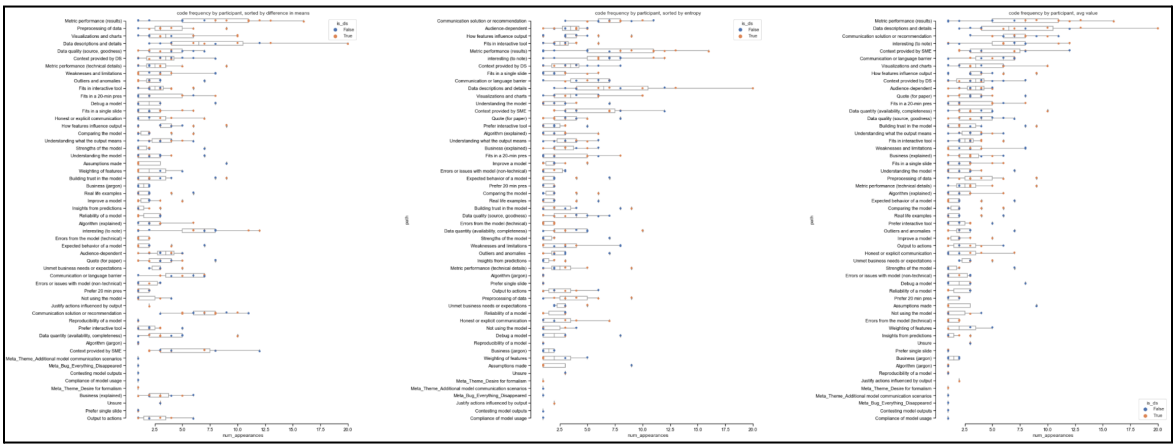
5. Authors 1 and 2 do a second pass by swapping each other's coded interviews over a two-week span. During the hand-off, authors 1 and 2 mark any disagreements they have in misused codes or missing codes from participant's utterances.

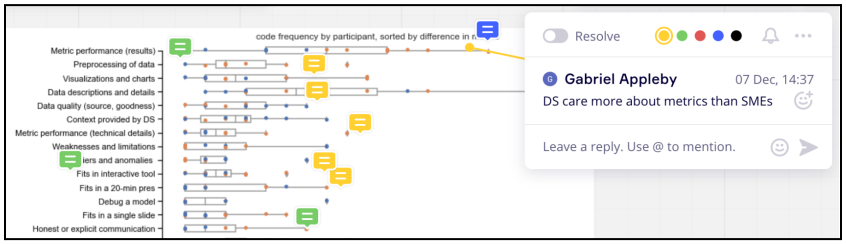
6. All authors meet to review disagreements in 4 total interviews (2 data scientists, 2 SMEs). All authors decide how general disagreements should be settled in code usage (i.e., particular discrimination & inclusion/exclusion criteria between code applications).

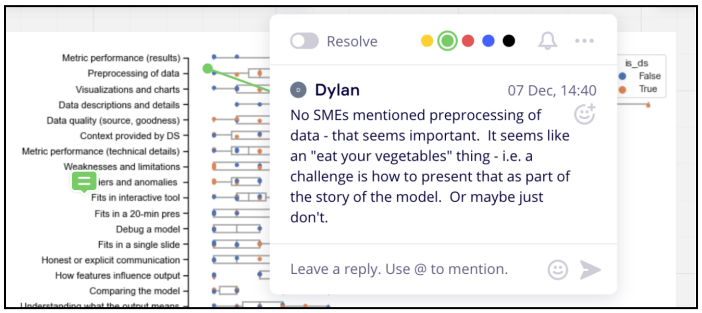
7. Authors 1 and 2 meet to review remaining disagreements from interviews. In total, 2 disagreements are leftover after discussion between authors 1 & 2. The final author (Author 6) makes the final tie-breaker.

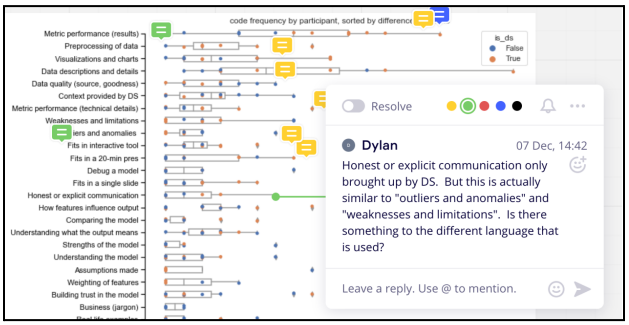
8. Authors 3 and 6 meet to take a final pass over all coded interviews.

4) **Analyzing frequency and difference of code usage between data scientists and SMEs**: After finalizing the codebook and coding all interviews, the authors held a brainstorming session to discuss differences in code frequency between data scientists and subject matter experts, using three different sorting methods for the code frequency: difference in means, entropy, and average value. Each author added comments to the charts to better understand why data scientists and SMEs seemed to value some model performance & model communication aspects more and/or differently (some examples of comments shown below).

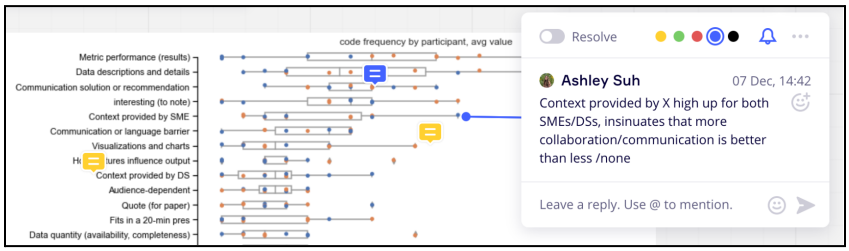


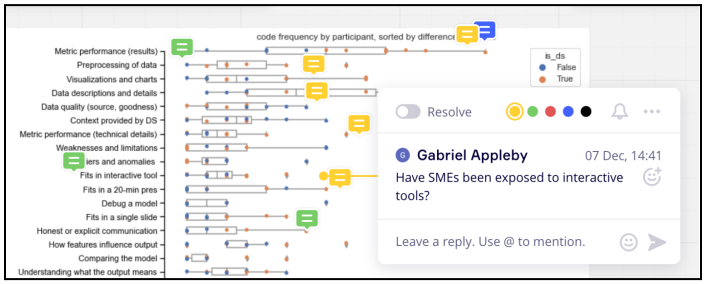


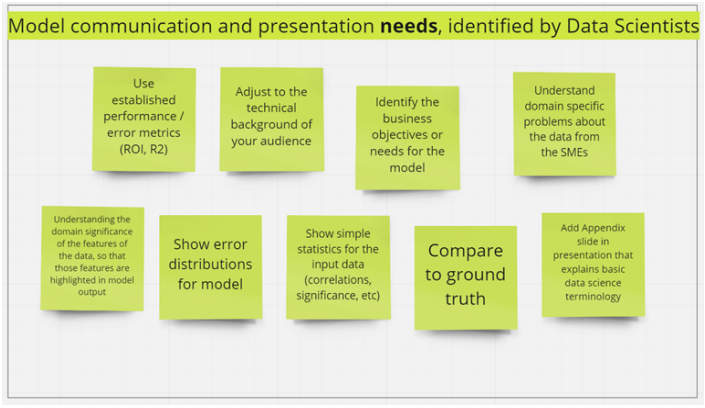


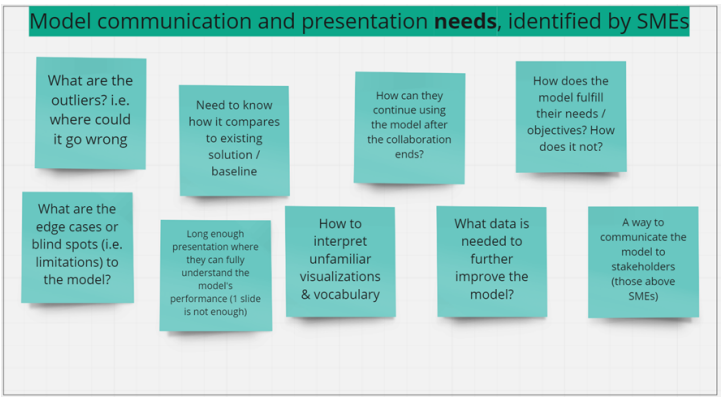


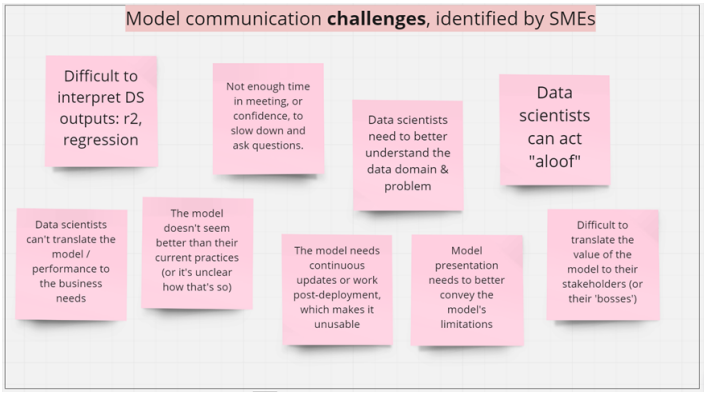



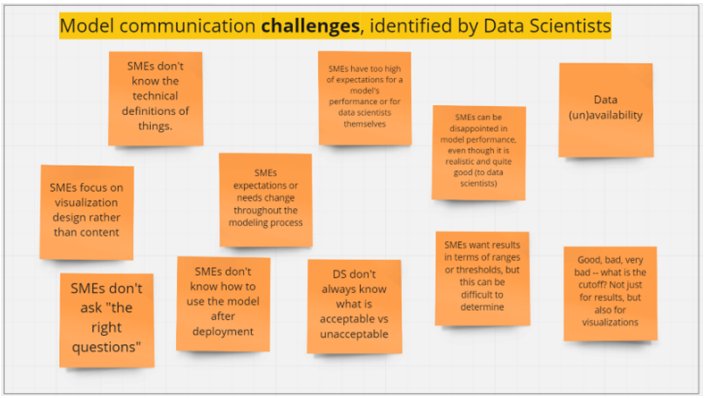
5) **Mapping interview findings to potential communication solutions**: The creation of our guidelines was an iterative process involving multiple brainstorming & design jam sessions with all authors. We describe each step below.

First, we identified the most important issues related to the needs and unaddressed challenges by data scientists and SMEs:

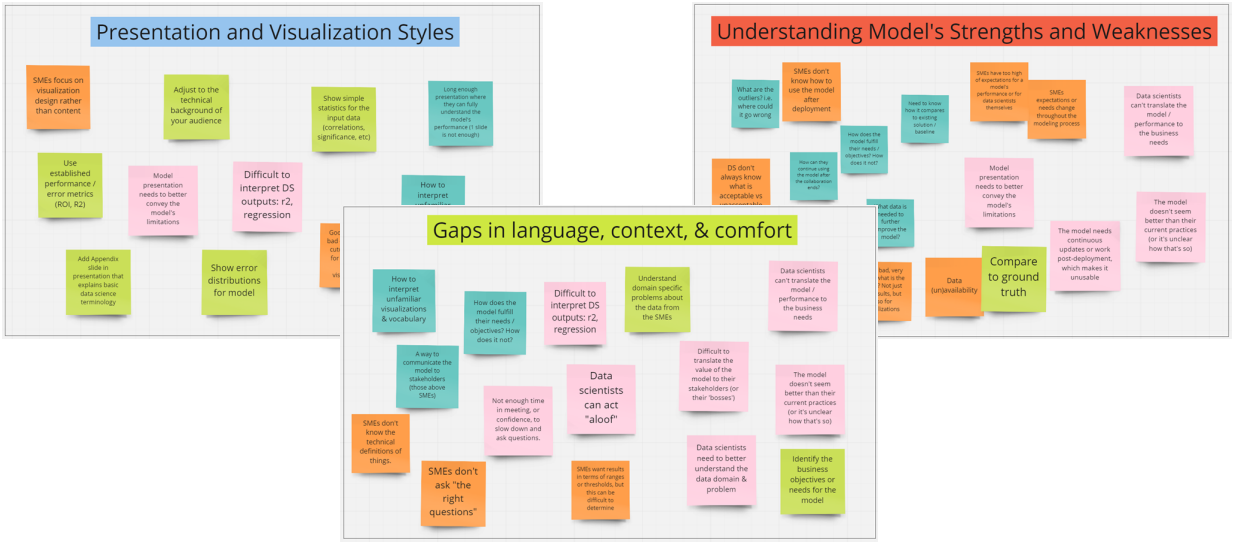


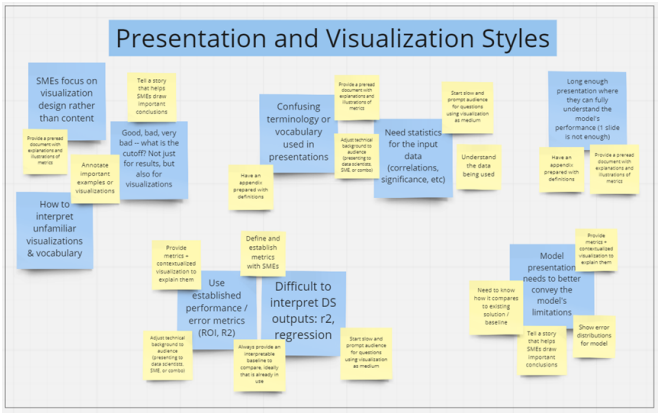



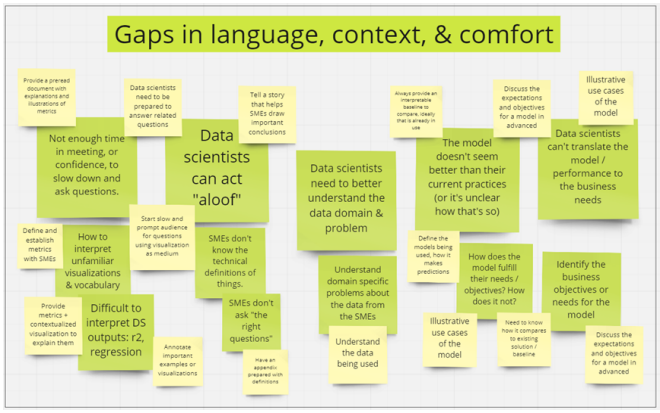



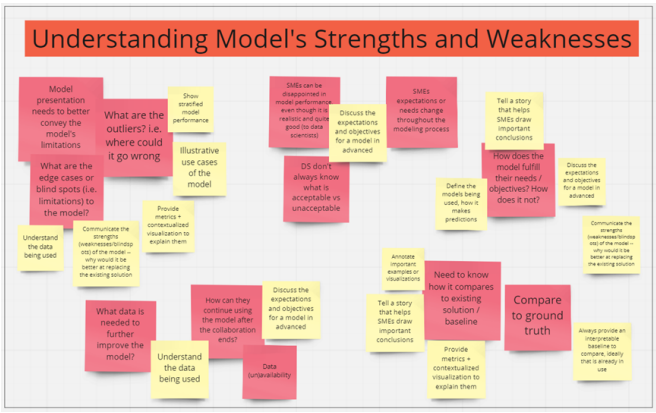

From these broad communication needs and unaddressed challenges, the authors decided on the 3 highest-level categories to encapsulate all individual issues: (1) Presentation and visualization styles; (2) Understanding the model's strengths and weaknesses; and (3) Gaps in language, context, & comfort. Each individual issue was mapped to its respective theme:



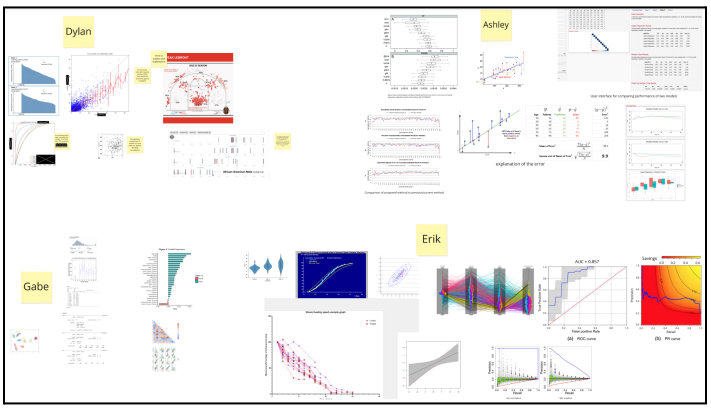
Because some issues could not be resolved to individual themes (e.g., a confusing visualization shown by data scientists can relate to `presentation & visualization style', as well as `gaps in language, context, & comfort'), authors considered the guidelines would have a many-to-many mapping from our interview findings (*communication challenges*) to our proposed solutions (*guidelines*). Therefore, we posed potential solutions to each of the issues for our three high-level themes. A solution could address multiple communication challenges, as so:

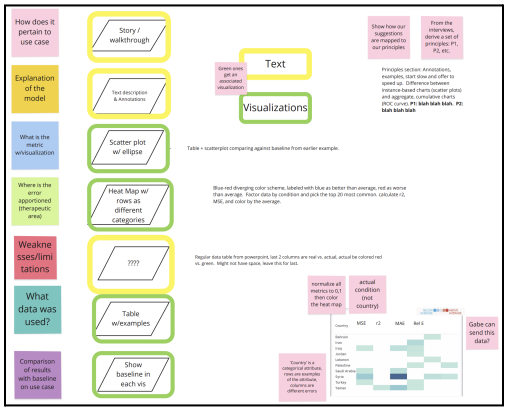



In addition to these communication solutions, the authors also posed potential visualization / visual communication solutions. Four of the authors spent roughly 15-20 minutes putting together screenshots from prior research contributions, novel ideas, and specific solutions brought up in our interviews. From these collective thoughts, we distilled them to generalizable visualization, tables, or text suggestions that could be mapped to a set of guidelines.



6) **Creation of the guidelines:** Finally, from the identified challenges from our interviews, the most applicable communication solutions that provided the most coverage to those challenges, in addition to our visualization and visual communication solutions, the authors decided on a final set of guidelines that could be broadly applicable and easy-to-implement for data scientists that must present data science/ML models to SMEs. For our paper, the guidelines were categorized into the same three categories from our high-level themes (presentation and visualization style, model strengths/weaknesses, and gaps in language/context/comfort).