Multimodal Technologies and Interaction

MDPI

Article

A Survey of Domain Knowledge Elicitation in Applied Machine Learning

Daniel Kerrigan [1,*], Jessica Hullman [2] and Enrico Bertini [1]

[1] Department of Computer Science and Engineering, New York University, Brooklyn, NY 11201, USA; enrico.bertini@nyu.edu
[2] Department of Computer Science, Northwestern University, Evanston, IL 60208, USA; jhullman@northwestern.edu
* Correspondence: djk525@nyu.edu

Abstract: Eliciting knowledge from domain experts can play an important role throughout the machine learning process, from correctly specifying the task to evaluating model results. However, knowledge elicitation is also fraught with challenges. In this work, we consider why and how machine learning researchers elicit knowledge from experts in the model development process. We develop a taxonomy to characterize elicitation approaches according to the *elicitation goal*, *elicitation target*, *elicitation process*, and *use of elicited knowledge*. We analyze the elicitation trends observed in 28 papers with this taxonomy and identify opportunities for adding rigor to these elicitation approaches. We suggest future directions for research in elicitation for machine learning by highlighting avenues for further exploration and drawing on what we can learn from elicitation research in other fields.

Keywords: elicitation; machine learning; domain expert; domain knowledge; expert knowledge

Citation: Kerrigan, D.; Hullman, J.; Bertini, E. A Survey of Domain Knowledge Elicitation in Applied Machine Learning. *Multimodal Technol. Interact.* **2021**, *5*, 73. https://doi.org/10.3390/mti5120073

Academic Editors: Alison M. Smith-Renner, Gonzalo Ramos and Gagan Bansal

Received: 16 October 2021
Accepted: 17 November 2021
Published: 24 November 2021

1. Introduction

Machine learning (ML) technologies are integrated into a diverse swathe of data-driven decision-making applications. Imbuing modeling with domain knowledge—declarative, procedural, and conditional information that a person possesses related to a particular domain—is a common goal. Expert opinion and judgment enter into the practice of statistical inference and decision-making in myriad ways across many domains. By obtaining and using expert knowledge, ML engineers or researchers can produce more robust, accurate, and trustworthy models.

Conducting expert knowledge elicitation in a way that ensures the gained knowledge will be useful can be challenging. ML engineers or researchers must target and associate domain knowledge with particular steps in the model development pipeline; for example, to guide problem specification, to inform feature engineering, and/or to aid with model evaluation. They must build common ground, mutual understanding of the context and targets of an interaction, with domain experts who may lack computational training. The elicited knowledge must then be integrated into the model, whether as an informal influence such as when domain experts' characterization of a problem is used to guide data curation, or as a more formalized influence such as when domain experts' labels or feedback are used directly to train or refine a model.

Despite growing research interest in the ways in which ML and AI systems interact with human knowledge and beliefs, why and how researchers are eliciting knowledge from domain experts in developing ML models has received little concerted attention. There are reasons to believe that eliciting knowledge from experts in a domain may call for different approaches and interfaces than one might use with non-experts such as crowdsourcing workers. Having more experience with a domain can change the mental representations one relies on [1] or ways in which one prefers to articulate their knowledge [2]. Fields like judgment and decision making, in which eliciting (often probabilistic) prior beliefs

is a topic of study, have provided evidence that how knowledge is elicited can affect the usefulness of the information. Some evidence of the effects of elicitation approaches can also be found in sub-areas of computer science that have focused mostly on non-experts like crowdsourced labeling (e.g., [3]) and active learning (e.g., [4]). However, to date few attempts have been made to characterize the space of decisions that ML researchers and practitioners make in eliciting domain knowledge from experts, and elicitation itself is rarely a topic in ML research.

One risk of a lack of awareness of elicitation practice is that many researchers are devising methods in an ad-hoc way, which may lead to less reliable elicited knowledge. As one scholar of expert elicitation for Bayesian applications has described, "[e]liciting expert knowledge carefully, and as scientifically as possible, is not simply a matter of sitting down with one or more experts and asking them to tell us what they think" [5]. Understanding the extent to which ML researchers and practitioners are anticipating and addressing challenges to elicitation, from possible cognitive bias to the need to establish common ground, to the value of using a systematic and reproducible elicitation method, can help point future research toward improved processes.

Our work characterizes ways in which researchers elicit domain knowledge for use in machine learning. Our goals in doing so are firstly to increase researchers' awareness of the variety of methods available, and secondly to foster discussion on where research may benefit from taking a more systematic or thoughtful approach. We survey elicitation practices used in 28 machine learning-themed research papers published between 1995 and 2020 [6–33].

Our first contribution is a taxonomy for characterizing domain expert knowledge elicitation approaches. At the highest level, our taxonomy distinguishes key considerations at four levels of decisions comprising knowledge elicitation. The first category, *elicitation goal*, captures the purpose of the elicitation by noting which part of the machine learning process the knowledge is used for. Second, *elicitation target* categorizes the kind of information that is elicited from the domain experts. Third, *elicitation process* examines the methodology of the elicitation, including the elicitation medium, the form of the prompts and responses, and the number of experts used. Lastly, *use of elicited knowledge* captures how the elicited knowledge is processed and incorporated in the machine learning process.

Our second contribution is an analysis of 73 "elicitation paths" that comprise distinct sets of choices related to eliciting knowledge across the 28 paper sample. We use these paths to reflect on common choices that emerge from the co-occurrence of codes applied to the research papers we surveyed. In particular, we analyze the different patterns that emerged for the four elicitation goals: *problem specification*, *feature engineering*, *model development*, and *model evaluation*.

Our final contribution is a set of recommendations based on the trends observed in the elicitation paths and insights from other elicitation literature. We seek to motivate an agenda for future research in elicitation for machine learning that includes increased emphasis on transparency and traceability in the elicitation process, establishment of common ground to support shared understanding between researchers and domain experts, addressing cognitive bias, and validating elicited knowledge, among others.

2. Related Work

2.1. Understanding ML Practice

Our work contributes to a growing line of research that seeks to understand machine learning practitioners' needs and practices via qualitative analysis [6,34–36]. Closest to our goals, several other studies have examined how data scientists and/or model developers collaborate with domain experts. Mao et al. [37] comment on the importance of, and challenges faced in, establishing and maintaining common ground between data scientists and biomedical scientists collaborating on data science projects, though did not focus on applied ML per se. In another interview study, Hong et al. [38] interviewed machine learning practitioners about model interpretability, finding that for many practitioners,

the primary value of interpretability was that it enabled more efficient communication between model developers and other stakeholders during model development and evaluation, many of which included domain experts. Other recent work examines needs of medical experts using AI-based tools through interviewing, finding that clinicians often desired basic global information about a model's properties, such as known strengths and limitations, over decision-specific information [39].

2.2. Knowledge Elicitation for Expert Decision Making

Studies of expertise reflect the characteristics, skills, and knowledge that distinguish experts from novices and less experienced people in various domains [40]. Research in cognitive and educational psychology has studied how gaining expertise in a domain can lead to differences in the content and organization of a person's knowledge [1,41,42]. As expertise became a topic of study in psychology, parallel developments in computer science in the late 1970s and early 1980s led to the study of expert systems in areas such as artificial intelligence and cognitive science, where expert knowledge needed to be elicited to try to replicate expert decision making processes [40]. Points of controversy concerning methods to elicit domain knowledge, either for study or system development, include whether experts can articulate the knowledge and methods that they use in complex situations to other people [40].

A related area of research concerned with elicitation is sometimes called the "classical" elicitation literature, which has generally been concerned with obtaining expert knowledge to construct a prior distribution for Bayesian statistical analysis [2,43]. While the question of whether the information one elicits from an expert is a perfect representation of their beliefs is hard to definitively answer, researchers have evaluated methods using properties such as how consistent responses are upon repeated queries as targets, or by providing a "ground-truth" distribution to experts to see how well it can be elicited back (e.g., [44]). Many established protocols for elicitation involve iterative collaboration between the statistician developing the model and the expert, as well as aggregation of multiple experts' beliefs where possible. Our work takes a closer look at elicitation practice in applied ML research for which some of the classical elicitation principles may hold, but may not be well-known among researchers or practitioners.

Some prior research surveys knowledge acquisition techniques for expert systems. In 2001, Wagner et al. [45] provided an overview of elicitation techniques including unstructured interviews, structured interviews, protocol analysis, psychological scaling, and card sorting. In more recent work, Wagner [46] analyzed 311 case studies of expert systems, finding that unstructured and structured interviews were by far the most common manual knowledge acquisition techniques, and that, over time, researchers tended to provide more detail about the elicitation process. Wagner noted, however, that a large number of papers (80) provided very little information about the process, for example making cursory references to talking to the domain expert only. In addition, Rahman et al. [47] recently reviewed the literature in databases, HCI, and visualization to describe a taxonomy for "amplifying domain expertise", or optimizing domain experts' interactions, through tools that enable the experts to interact more meaningfully at all stages of the pipeline for data-driven decision support. Our work is similar in spirit to these studies, but we focus on expert knowledge acquisition in the context of ML model development over the simpler rule-based approaches that dominate the expert systems literature, and we focus more on methods than trends based on, for instance, application areas.

Research in crowdsourcing and various forms of human-in-the-loop machine learning, such as active and interactive learning and machine teaching, have taken structured approaches to eliciting knowledge and have commented on their effects and challenges. Active and interactive machine learning are areas where human knowledge is used to support data labeling and the specific task of model refinement through label corrections, a combination of model development and evaluation. Though these areas may use non-expert knowledge, some lessons from this research are informative for elicitation of domain

knowledge more broadly. For example, in summarizing challenges and opportunities in interactive machine learning, Amershi et al. [48] describe how active learning can lead to users being overwhelmed by questions [49], how people can tend to give more positive feedback which sometimes harms learning [50], and the value of providing contextual information to support label quality [51]. Close to our goal of outlining a research agenda, they motivate the need for richer interfaces both for eliciting input from humans and presenting model output. Such interfaces can be used for tasks that go beyond labeling to include relevant tasks such as feature creation, re-weighting features, adjusting cost matrices, or otherwise modifying model parameters. Others have pointed to specific threats in interactive ML related to the information a human provides, including that a user's input reinforces noise in the training data or statistics they see [52]. Recent works have also surveyed research in human-in-the-loop machine learning [53,54]. Some of the papers covered in our survey propose human-in-the-loop systems. In these cases, we are specifically interested in how they elicit and use knowledge from domain experts for the purpose of creating better machine learning models.

Crowdsourcing studies aimed at eliciting labels for training datasets from non-experts have also shown how differences in how information is shown [3] or what sorts of interruptions a crowdworker experiences [55] can impact the quality of labels obtained. Others have explored the value of incentivizing label and other data collection from non-experts according to their usefulness for model development (e.g., [56]).

3. Materials and Methods

The goal of our analysis is to identify how ML researchers are making decisions about domain knowledge elicitation. To do so, we collected a set of research papers published between 1995 and 2020 that describe or motivate the elicitation of domain knowledge for the development of machine learning models. We describe the scope of our analysis, sample collection, and coding process.

3.1. Scope

Our interest is in the explicit elicitation of domain expertise to improve ML models. We include both papers where researchers apply ML to a particular domain and elicit knowledge from experts and papers that present ML-related systems, tools, or algorithms that involve expert elicitation. While many research articles on ML-advised decision making may mention the importance of human knowledge, given potential differences between how domain experts versus novices represent and articulate their knowledge, we focused on research aimed at eliciting expertise that was implied to be held by certain groups of professionals, and hence could not be easily obtained from the online or university student recruitment pools.

We define explicit elicitation as elicitation of knowledge where the domain expert is aware that they are providing input, and where the elicitation process is implied to occur as part of the course of research described in the paper. This definition eliminates, for instance, papers that describe using pre-existing domain knowledge, such as existing knowledge bases, to develop or refine an ML model. It also excludes systems that rely on completely implicit elicitation. Examples include systems with embedded learning components that improve the system using the expert's interactions but without the expert being aware of any elicitation, as well as implicit use of domain knowledge that occurs when one of the researchers tightly involved in the ML research is a domain expert. At least seven of the papers included in our analysis had a domain expert listed as an author [7–13]. In such cases we coded for any described elicitation, but our taxonomy does not cover implicit use of domain knowledge that may have also affected these projects.

Finally, our interest in domain knowledge elicitation for the purpose of producing a better ML model than would be possible without it precludes the inclusion of ML research where experts are simply used to evaluate a system or pipeline, with no intention of using the elicited knowledge to improve the system or pipeline that the paper presents. Likewise,

research where domain experts use ML models without explicitly providing knowledge for the development or improvement of the models does not meet our criteria.

3.2. Sample Collection

Our search for papers to include in our analysis began with Google Scholar searches for combinations of terms such as "domain experts", "elicitation", "data science", and "machine learning". These searches returned a large number of papers, many of which were not specific to machine learning, expert elicitation, or other constraints described above. We added to our search-based samples papers that we obtained by following relevant references cited by these works. The prevalence of ML-advised decision making in an increasingly diverse set of domains led us to aim for a representative, but not necessarily complete sample. We therefore next performed additional searches for domains that we did not feel were adequately covered, such as in using ML for healthcare and for expert labeling tasks. We also solicited papers on social media, requesting "ML/human in the loop research papers that describe eliciting knowledge from domain experts". Each of these methods turned up a number of papers that met some, but not all of our criteria for explicit elicitation of knowledge from domain experts. We report our analysis on 28 papers that passed our criteria, collected over roughly six months of iterative search and coding. We closely read approximately 50 total papers to determine their fit with our criteria, since many papers talked about domain knowledge or elicitation but needed to be read to determine whether or not they involved any actual elicitation of expert knowledge.

3.3. Content Analysis

Our goal in analyzing the papers was to learn about how eliciting knowledge from domain experts is performed and incorporated into the machine learning process. We made iterative qualitative coding passes through the papers, using standard open coding procedures [57]. We started by pulling out relevant details about working with domain experts for open coding and taking note of the main aspects to the elicitation approaches that emerged. We sought a consistent set of dimensions that could be used to break down and categorize a given elicitation methodology. Each time we added a new paper to our sample that introduced a new method or goal of elicitation, we updated our taxonomy and recoded prior papers as needed. Coding was led by the first author, but ambiguous methods or otherwise difficult judgments were discussed by all three authors during weekly meetings over the course of a roughly six-month period. We present the resulting taxonomy in the next section.

Within each paper, we identified one or more "elicitation paths" that represented unique combinations of the elicitation goal, intended target, elicitation process details, and descriptions of how elicited knowledge was used in the ML pipeline. For example, if a paper presents that the same elicited information was used for two different purposes, then the uses would be listed as separate paths. Likewise, if the same methodology, such as interviews, was used to elicit two different categories of information, then they would be listed separately as well. In the 28 papers that we analyzed, we identified a total of 73 elicitation paths. The Supplementary Materials include a spreadsheet that contains the list of analyzed papers, the taxonomy hierarchy, and the codings for all elicitation paths.

4. Elicitation Taxonomy

The top level of our taxonomy consists of four high level categories: *elicitation goal*, *elicitation target*, *elicitation process*, and *use of elicited knowledge*. Each of these are further divided into subcategories. After the name of each code, we list in parentheses the number of times that code appeared and the percentage of paths that contain it. Figure 1 gives a visual overview of the coding of the 73 elicitation paths according to the taxonomy. The Supplementary Materials include the code to generate the visualizations in this work.

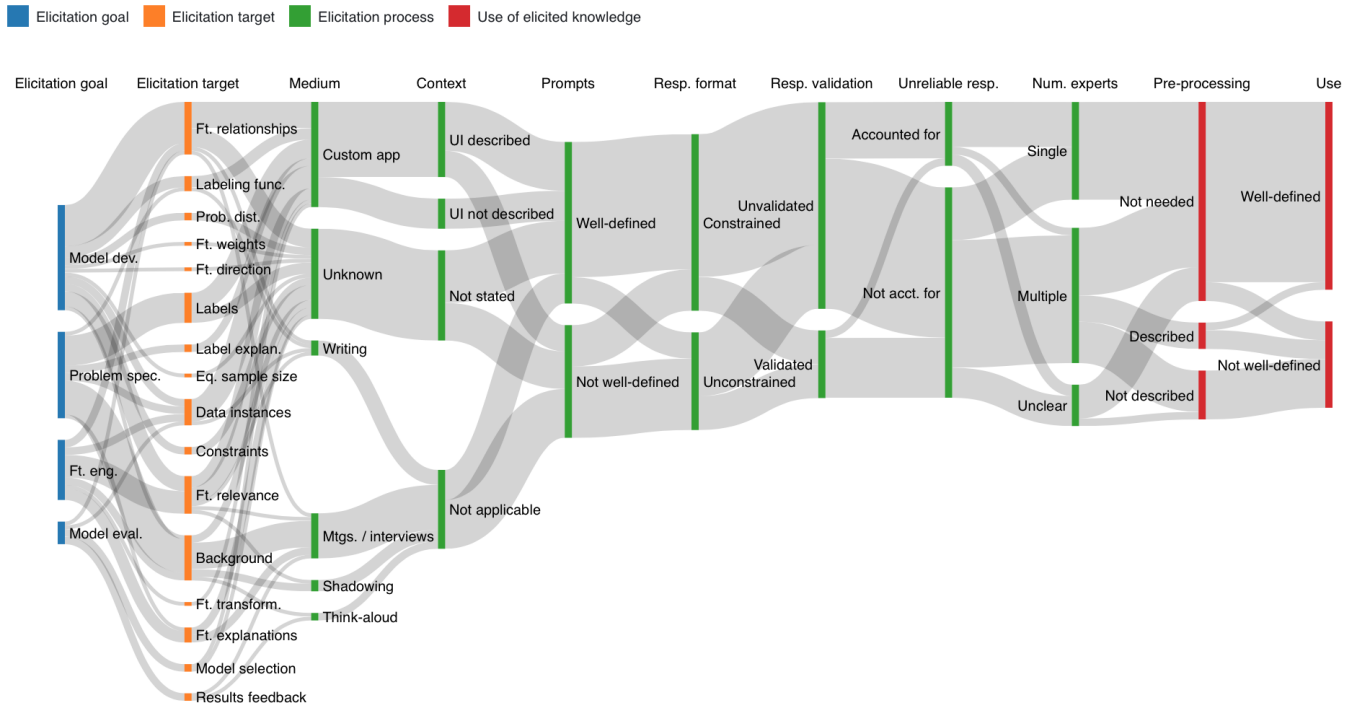

Figure 1. This Sankey diagram shows the 73 elicitation paths coded according to our taxonomy. Each node represents one low-level code in the taxonomy. The color of a node encodes the top-level category that the node belongs to. The horizontal position of a node encodes the middle-level category that the node is under.

4.1. Elicitation Goal

Elicitation goal describes different ways that knowledge from domain experts is implied to be useful in the model building process. The ML and related literature demonstrates at least four areas where incorporating expert knowledge can be useful. The first is *problem specification* (23/73, 32% paths), which includes defining the task the model should solve, understanding the domain experts' current practices for the task, identifying metrics to use for evaluation, and gathering training and testing data. Second is *feature engineering* (16/73, 22%), which includes elicitation of feature level information used to determine what the inputs (i.e., features) to the model should be. Feature engineering concerns the format of the data instances that are input to the model. Third is *model development* (28/73, 38%), the most common goal of elicitation in our sample, which includes defining the model structure and parameters. We separate feature engineering from model development based on a common distinction in machine learning terminology between features as inputs to a model and model parameters representing configuration variables or hyperparameters internal to the model, as well as choices related to a model's structure. The distinction between the two can be subtle. For example, while in many cases, eliciting feature information falls under feature engineering, if that information affects internal model parameters or configuration, eliciting feature information can be better described as targeting model development. Fourth is *model evaluation* (6/73, 8%), which includes assessing the model's performance and validating its results for the purpose of improving the model.

4.2. Elicitation Target

Elicitation target refers to what form of knowledge the elicitation is intended to obtain from the domain experts. *Background knowledge and processes* (12/73, 16%) includes information about the domain, the task at hand, the needs of the domain experts, and their workflows. *Labeling functions* (4/73, 5%) are user-specified functions that capture the logic used to decide which label should be assigned to an instance, such as if–then rules or heuristics. *Labels* (8/73, 11%) are assignments of classifications to individual instances.

Label explanations (2/73, 3%) are comments provided by the expert discussing their reasoning for choosing a particular label for a given instance. *Data instances* (7/73, 10%) are specific data samples or data points elicited from the domain experts, such as such as edge cases or other examples of cases of interest. *Probability distributions* (2/73, 3%) include prior distributions for a Bayesian network or probabilistic labels. *Feature explanations* (4/73, 5%) are definitions for features, such as what a feature means and how it is derived. *Feature relationships* (14/73, 19%) outline how features are connected to each other, such as causal relationships or hierarchical relationships. *Feature relevance* (10/73, 14%) indicates labels of whether or not a feature is important for the task, or lists of features that are or are not deemed important. *Feature transformations* (1/73, 1%) are functions chosen by the domain expert to pre-process features used by the model. *Feature weights* (1/73, 1%) are elicited values that indicate the magnitude of feature impacts on the outcome. *Feature direction* (1/73, 1%) notes whether a feature has a positive, negative, or no effect on the outcome. *Equivalent sample size* (1/73, 1%), as defined by Heckerman et al. [14], is the size of a dataset that someone starting from ignorance would need to see in order to be as confident in their prediction as the domain expert. *Model constraints* (2/73, 3%) are limits on values that variables can take or on how they can behave, such as monotonicity constraints. *Model selection* (2/73, 3%) refers to a domain expert's choice of a specific trained model from a series of alternatives. *Results feedback* (2/73, 3%) covers the domain expert assessing the output of a model, such as working with the domain expert to iteratively refine the model's classification threshold and validate the model's output [11] or receiving feedback from domain experts that model behavior does not match their expectation in a specific situation [15].

4.3. Elicitation Process

Elicitation process characterizes how knowledge is elicited from the domain experts. If we view elicitation as a "conversation" between the domain expert(s) and the ML researcher or ML system, then we can ask about the medium that the conversation occurs through, who it is between, how it is structured, and how common ground is ensured. First, we look at the *elicitation medium*, which captures how this conversation takes place and what the domain experts use to communicate their knowledge. Next, we consider the *context* that the domain expert receives during elicitation. Effective conversation requires establishing common ground [58], so we consider the extent to which the information given to the domain expert about the model or the eliciter's goals is described. Following this, we cover the *number of domain experts* that knowledge is elicited from. We also categorize how structured the *prompts* that the domain experts receive are and how constrained their *responses* are, as well as whether or not the responses are *validated* and how *unreliable responses* are handled.

The medium of elicitation describes the type of communication channel through which the information is obtained. *Custom app* (28/73, 38%) refers to a computer application made specifically for the given task. *Writing* (4/73, 5%) covers written communication, such as with pen and paper or writing in a text editor or other application, for example. *Meetings or interviews* (12/73, 16%) includes verbal communication through in-person and remote meetings, interviews, and workshops. *Shadowing* (3/73, 4%) occurs when the researcher is observing the domain expert perform their work. In a *think-aloud* (2/73, 3%), the domain expert shares their thought process while performing a task or using a tool. If a path does not describe how the elicitation is meant to be performed, then it falls under the *unknown* (24/73, 33%) category.

We categorize approaches to providing *context* to domain experts into cases where the elicitation is performed through a custom app under *UI described* (20/73, 27%) or *UI not described* (8/73, 11%), depending on whether or not the user interface of the application's elicitation component for the path is shown or detailed in the paper. In the case where the elicitation medium was unknown, then the context is categorized as *not stated* (24/73, 33%).

If the elicitation is known to not be done through a computer application, then its context falls under *not applicable* (21/73, 29%).

Next, we consider the *prompts* that the domain expert receives and the *responses* that they give during elicitation. There are *well-defined prompts* (43/73, 59%) if the process is described in a way that implies that the expert is responding to clearly specified questions or tasks. For example, we surmise that a structured interview or labeling task will have well-defined prompts, whereas an elicitation process described as a casual meeting or as an open-ended conversation would have prompts that are *not well-defined* (30/73, 41%). Related to the degree of systematicity in the prompts, the response format is *constrained* (47/73, 64%) if the domain expert is restricted in the information that they can provide or the actions that they can perform. These constraints could include having the expert choose from a pre-defined set of choices or restricting the responses to a particular format, such as requiring a number. For example, the domain expert being limited to providing if–then rules would be constrained, but free-form responses, such as from a meeting, would be *unconstrained* (26/73, 36%).

We consider the *number of experts* that researchers or the system elicits information from, either *single* (26/73, 36%), *multiple* (36/73, 49%), or *unclear* (11/73, 15%). If a path is evaluated with a user study involving multiple experts and the responses from the experts are used and analyzed individually, then it is classified as *single*.

We also consider if there is any validation on the information provided by the domain expert. A response is *validated* (18/73, 25%) if it is checked for correctness in some way, even weakly, such as by using multiple domain experts and having an explicit strategy to handle disagreements, for example through majority vote or having them jointly reach a consensus. If there is no mention of checking the information provided by the domain expert, then we categorize the responses as *unvalidated* (55/73, 75%). In addition, we consider if unreliable responses are explicitly *accounted for* (17/73, 23%) or *not accounted for* (56/73, 77%) in using the information from the expert. For example, a learning algorithm may factor in the probability that the expert is incorrect. Accounting for unreliable responses is distinct from validating responses, as validation is an attempt to check the correctness of elicited information before it is used, whereas accounting for unreliable responses tends to occur in the procedure to integrate the expert's information into the model. For example, one approach might have one expert verify the information that was elicited from another expert and then assume that the information is correct for the rest of the process. Another approach might elicit from a single expert, not validate the responses, and then explicitly account for that information being possibly incorrect when using it.

4.4. Use of Elicited Knowledge

We first look at whether manual pre-processing or analysis is needed on the elicited information before it is used. Pre-processing is not needed when the elicited information is directly used, such as when it is a direct input to the model or another algorithm. In other cases, pre-processing or analysis is needed before using the information, such as using grounded theory to analyze interviews and observations [11], formalizing expert knowledge into if–then rules [16] or flowcharts [17], and standardizing feature definitions obtained from different sources [12]. We therefore categorize pre-processing as being *not needed* (53/73, 73%), *described* (7/73, 10%), and *not described* (13/73, 18%). If there is not a way to directly use the elicited information and there is no detail given about how the elicited information was processed or analyzed, then we categorize it as *not described*.

Finally, we classify the use of the information as either *well-defined* (50/73, 68%) or *not well-defined* (23/73, 32%). We consider elicited information to have a well defined use if there is a predetermined, unambiguous, or algorithmic way to incorporate it in the machine learning process. For example, eliciting labels, monotonicity constraints, or whether or not there should be an edge between two nodes of a Bayesian network can all have clear and well-defined uses. This is in contrast to the experts' background knowledge, current

process, and general feedback on model results, which might be less directly actionable and pose more degrees of freedom from the researcher in how they use them.

Figure 2 visualizes counts of the number of elicitation paths for each code described above, separated by the path's goal.

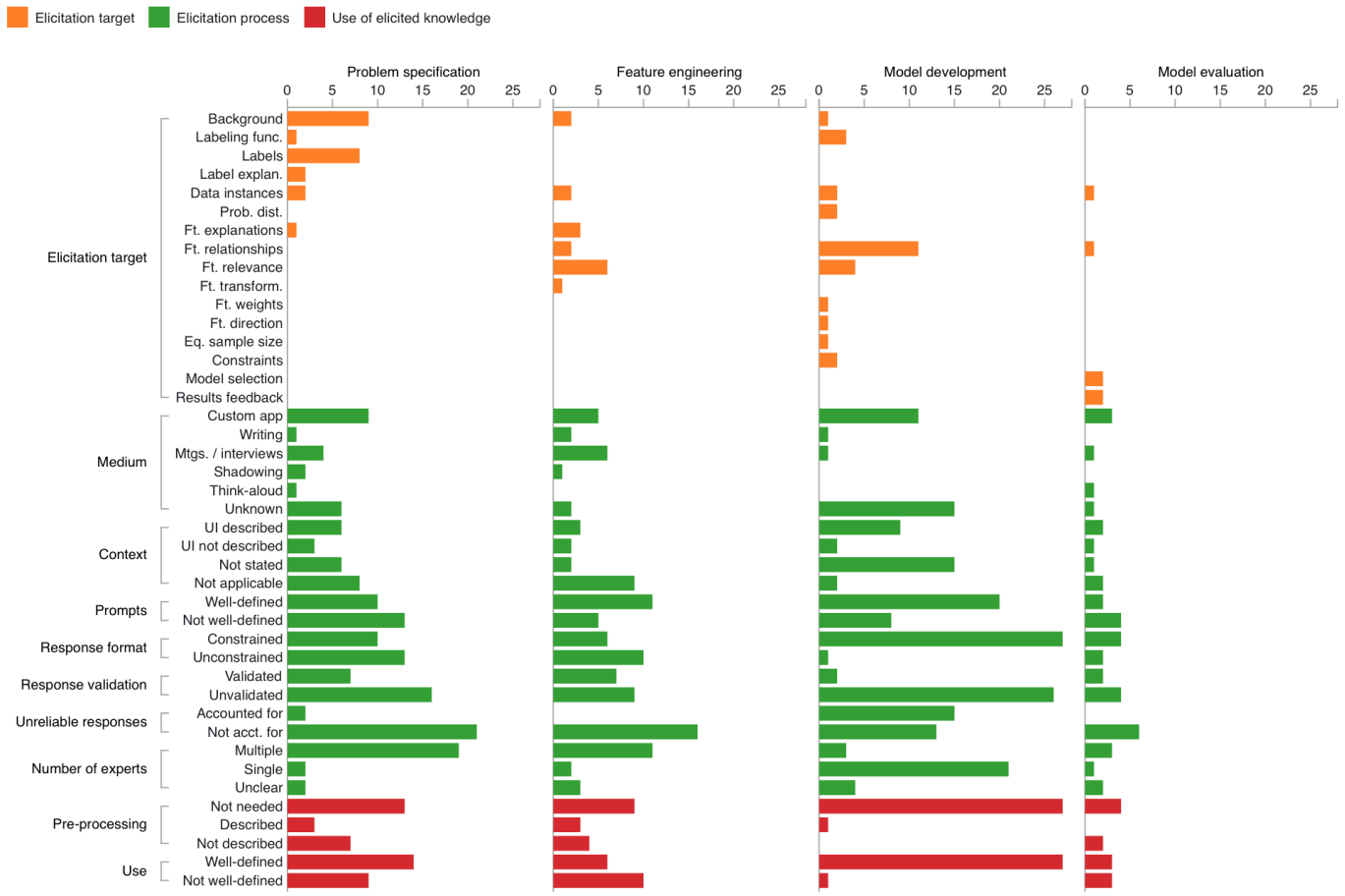

Figure 2. These bar charts show the number of times each low-level code in our taxonomy appeared in the elicitation paths, broken down by the elicitation goal.

5. Results

We describe observed differences in elicitation paths based on their goals, and point to gaps and opportunities for knowledge elicitation in ML that emerge from our analysis.

5.1. Characterizing Elicitation Paths

By analyzing how the taxonomy codes co-occur in the elicitation paths, we can identify common elicitation trends. We describe observations from separating the paths according to their elicitation goal.

5.1.1. Problem Specification

Problem specification accounted for 23 out of the 73 total paths (32%), which are visualized in Figure 3. These paths were mostly split between eliciting *background knowledge and processes* (9/23, 39%) and eliciting *labels* (8/23, 35%). *Feature explanations* and *labeling functions* each appeared a single time and *data instances* and *label explanations* appeared twice, all for the preparation and collection of training and testing data.

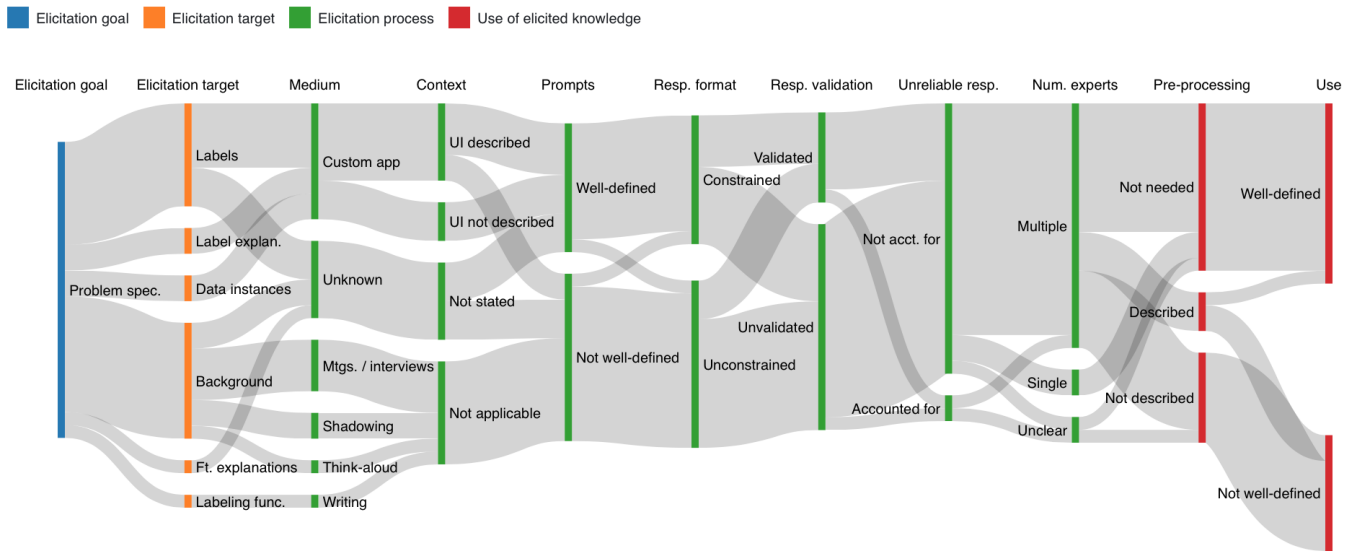

Figure 3. This Sankey diagram shows the 23 elicitation paths for problem specification.

Background knowledge and processes were most commonly elicited via *meetings or interviews* (4/9, 44%), followed by *shadowing* (2/9, 22%), *unknown* (2/9, 22%), and *think-alouds* (1/9, 11%). In all cases, the prompts were *not-well defined*, the response formats were *unconstrained*, and unreliable responses were *not accounted for*. In the majority of paths, responses were unvalidated (7/9, 78%) and *multiple* experts were used (8/9, 89%). Two paths *described* how the elicited knowledge was pre-processed or analyzed before it was used.

The media for eliciting labels were *custom app* (5/8, 63%) or *unknown* (3/8, 38%). Two of the *custom apps* that elicited *labels* also elicited *label explanations*, in which experts justified and discussed their labels in order to reach a consensus label [13,18]. Other forms of label validation in order to resolve disagreements between experts included using majority vote [9,13] or consulting a more senior expert [10]. A few paths that elicited labels did not explicitly aggregate the results of multiple experts to detect and resolve disagreements [11,15,16,18]. In all paths that elicited labels, the response format is *constrained*, manual pre-processing is *not needed* before using the labels, and their use is *well-defined*.

5.1.2. Feature Engineering

Feature engineering appeared in 16 out of 73 paths (22%) geared towards determining feature inputs to a model. These paths are shown in Figure 4. *Feature relevance* (6/16, 38%) was the most frequent elicitation target, followed by *feature explanations* occurring three times, *data instances* (used to identify which features were important to capture), *feature relationships*, and *background knowledge and processes* appearing twice, and *feature transformations* showing up once. *Meetings or interviews* (6/16, 38%) and *custom apps* (5/16, 31%) were the most common media. The majority of paths had *well-defined* prompts (11/16, 69%) and elicited knowledge from *multiple* experts (11/16, 69%). Unreliable responses were *not accounted for* by any of the paths, and most of the uses of the elicited knowledge were *not well-defined* (10/16, 63%), suggesting the researchers' discretion was used to determine how much weight to place on the experts' advice about features. The 16 paths appeared across 7 papers, with 4 papers eliciting multiple types of information for feature engineering.

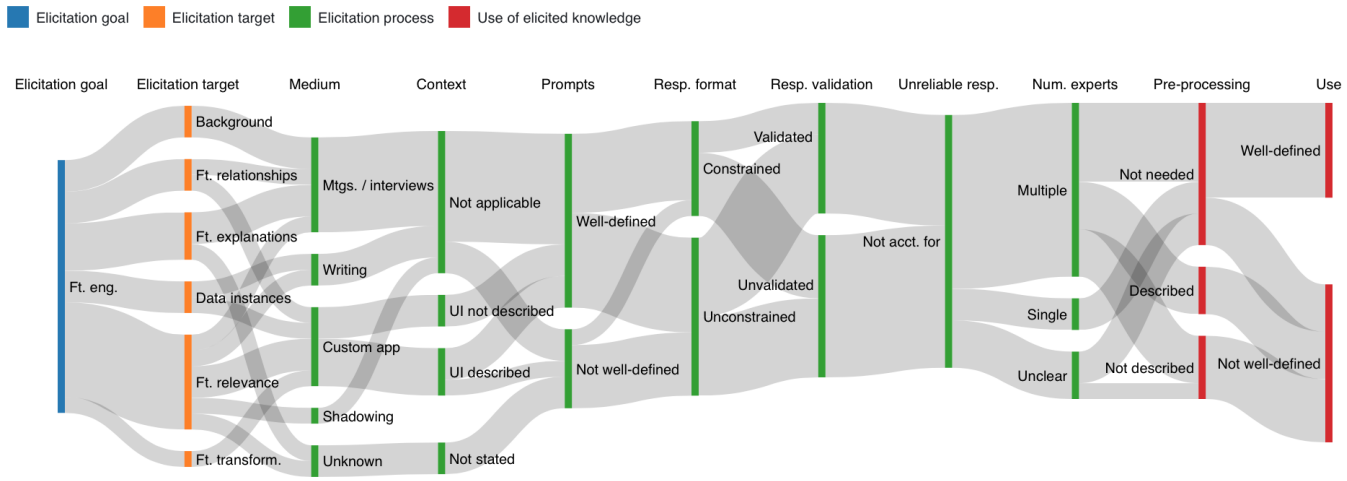

Figure 4. This Sankey diagram shows the 16 elicitation paths for feature engineering.

5.1.3. Model Development

Model development (28/73, 38%) was the most common elicitation goal. These paths are visualized in Figure 5. The highest frequency elicitation target for these paths was *feature relationships* (11/28, 39%), followed by *feature relevance* (4/28, 14%). In contrast to these elicitation targets as they appear under feature engineering, in model development the relationships or relevance statements directly affect model parameter estimates or other aspects of model structure. For example, one of the more common forms of *feature relationships* was eliciting the existence of edges in a Bayesian network, with some paths querying the expert about the existence of specific edges between pairs of variables [19,20] and others asking for information about all edges in the network [14,21,22]. Other forms of feature relationships include if pairs of features have the same or different directional impact on the outcome variable [23] and the relationships between words, topics, and documents in topic modeling [24,25].

Many of these paths were focused on the algorithmic aspects of how to use the elicited knowledge in the model. Often the elicited knowledge was a direct input to an algorithm. Given this, we see that in nearly all paths the responses are *constrained*, manual pre-processing or analysis is *not needed*, and the use of knowledge is *well-defined*. The one path with unconstrained responses and a need for pre-processing comes from Lee et al. [16], where domain experts were interviewed for feature-based knowledge of how to classify instances, which was then formalized into if–then rules for a rule-based model. The emphasis on the algorithmic uses of elicited knowledge in these paths may have come at the expense of attention paid to the human-centered aspects of how the knowledge is elicited. For example, the elicitation medium is *unknown* (15/28, 54%) for a majority of the paths.

These paths also stand out in that eliciting knowledge from a *single* expert was much more common than eliciting knowledge from *multiple* experts. This contrasts with the groups for *problem specification* and *feature engineering*, which were dominated by paths eliciting from *multiple* experts. In addition, these paths have unreliable responses *accounted for* (15/28, 54%) at the highest rate. The *model development* paths represent all but two of the appearances of this code across all paths. This is often done by modeling the probability of the elicited knowledge being incorrect or modeling the expert's uncertainty [7,14,19–22,26].

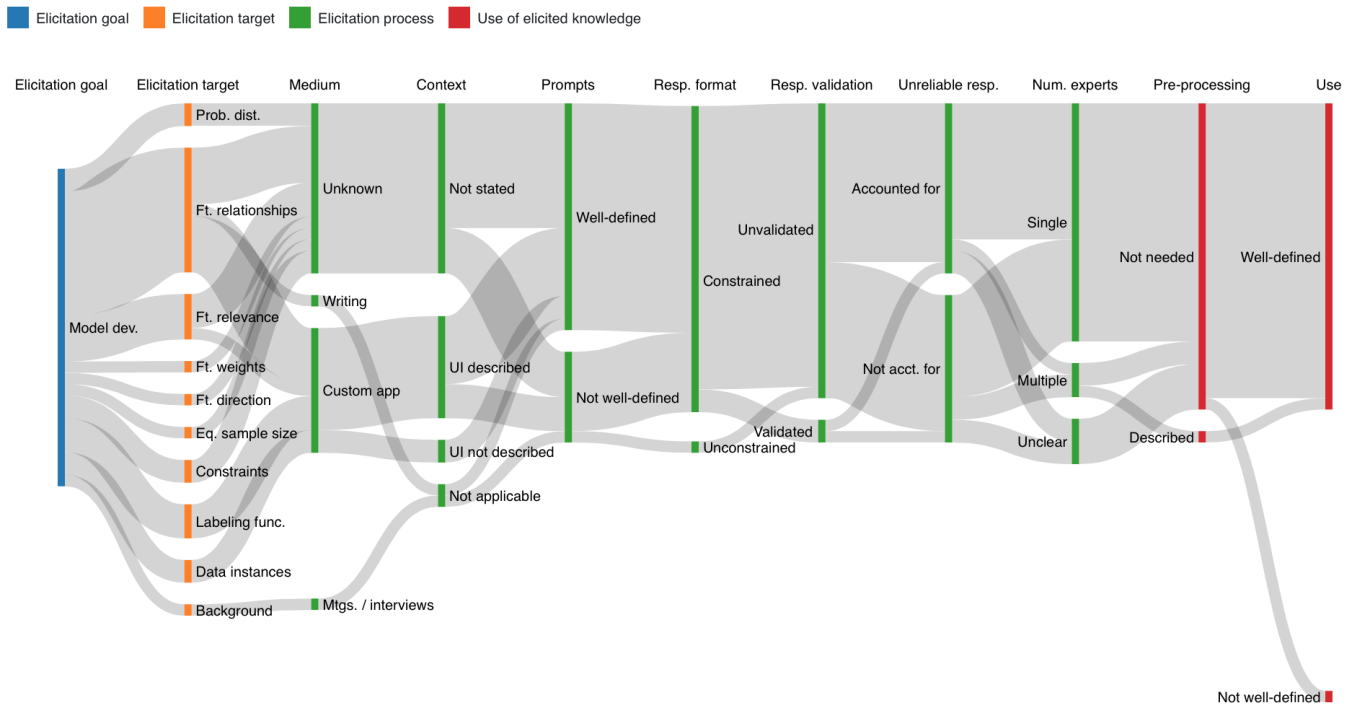

Figure 5. This Sankey diagram shows the 28 elicitation paths for model development.

5.1.4. Model Evaluation

Model evaluation (6/73, 8%) was the least common elicitation goal. These six paths are shown in Figure 6. We observed two occurrences each of *model selection* and *results feedback* and a single occurrence of *feature relationships* and *data instances* as the elicitation target. A *custom app* was the most common medium and two out of those three had their *UI described*. The majority of paths had prompts that were *not well-defined* (4/6, 67%). The use of the elicited knowledge was an even mix of *well-defined* (3/6, 50%) and *not-well defined* (3/6, 50%).

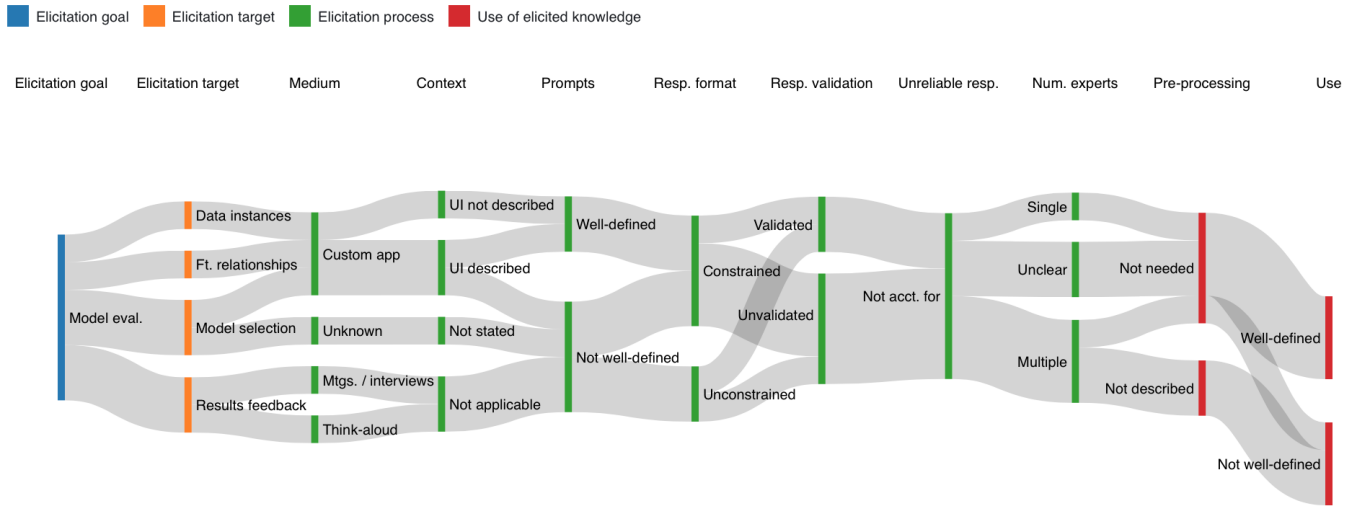

Figure 6. This Sankey diagram shows the 6 elicitation paths for model evaluation.

5.2. Gaps and Opportunities

Our analysis led us to observe several ways in which researchers are not necessarily treating elicitation as systematically or scientifically as they might. We list these below as opportunities for enhancing practice. Where relevant, we draw on principles from knowledge elicitation research outside of the machine learning literature, including the well developed area of expert prior elicitation for uncertain quantities [2,5,43,59]. While eliciting *probability distributions* only represented 2/73 elicitation paths in our work, eliciting information about uncertain phenomena that experts have extensive real world experience with makes the lessons of this literature applicable in many cases.

5.2.1. Transparency and Traceability

A total of 24/73 (33%) of elicitation paths had an *unknown* medium, meaning they did not contain information on how the knowledge is elicited. An amount of 8 out of the 28 paths that used *custom apps* did not describe the user interface for elicitation. In addition, *meetings or interviews* appeared in 12/73 paths (16%), but only 5 had *well-defined* prompts. Such a lack of detail highlights what may be a tendency to overlook the importance of carefully designing the elicitation process to ensure valid results. Clear documentation of elicitation practices is a principle in much of the expert elicitation literature outside machine learning due to its ability to provide both *traceability* (making it possible for those engaged in model development to more easily identify sources of various information that has been used in model building, for example, for debugging purposes) and *transparency*. Particular elicitation protocols have been developed and evaluated, such as the SHELF [60,61], Cooke [62], and Delphi [63] protocols, and provide more systematic and careful approaches to expert knowledge elicitation than many of the paths that we have covered in this work. For example, the SHELF protocol uses predefined templates to guide the elicitation's correct execution and to document the process, adding transparency and traceability [5,59]. Bowles et al. [13] was the only paper in our survey that explicitly mentioned using one such protocol by name. In their case, they used Delphi to resolve labeling disagreements among multiple experts.

There are also potential benefits to documenting the elicitation process and the use of the elicited knowledge so that the domain experts can see how the information that they provide is used, potentially leading to greater trust and opportunities for them to correct misinterpretations. The importance of transparency when using domain knowledge for ML has also been noted in Mao et al.'s interview study [37] and Rahman et al.'s survey, which states that the "critical and subjective nature of these [medical] decisions necessitates transparency, both from the algorithm as well as domain experts". For elicitation paths that used *meetings or interviews*, the workshop-based elicitation methodology proposed by Seymoens et al. [17] stands out in this regard, given its systematic and documented approach. Three of five paths that used *meetings or interviews* with *well-defined prompts* came from Seymoens et al. [17] and the other two from Hu et al. [27], who provided the list of questions asked during their interviews.

5.2.2. Systematic Use of Elicited Knowledge

There were 20 paths where the elicited knowledge could not be directly used and therefore manual pre-processing or analysis was needed before incorporating the knowledge in the machine learning pipeline. Of these paths, 7 *described* the manual processing or analysis that was performed, such as how the information was coded, formalized, or standardized, for example. This is opposed to 13 of these paths where it was *not described*. In addition, in 23 out of 73 paths, the use of the elicited knowledge was *not well-defined*, meaning that there was a not a clear, unambiguous way to use the knowledge provided by the expert. In some cases, this may be necessary, such as with feature engineering, where the ML researcher may need some degrees of freedom in order to translate the information provided by the domain expert into concrete features. However, situations where the elicited knowledge cannot be used without some processing, the pre-processing or analysis

is *not described*, and the use is *not well-defined* can pose risks to the reproducibility of the methods used, which may directly affect the impact of the research.

5.2.3. Motivating What Is Elicited

While we did not code specifically for how researchers motivated why they elicited the information they did from experts, we note that there is an opportunity for researchers to more deeply discuss the choice of what to elicit based on user psychology. In the classic prior elicitation literature, significant research attention has been paid to understanding how people best relate to the phenomena or concept that is the topic of elicitation, such as probabilities of events. For example, one way that this has manifested is in prioritization of frequency-based framings of probability in elicitation, to better align with people's seeming proclivity for reasoning using frequency over continuous representations of probability (e.g., [44,64]).

Machine learning researchers may also find it useful to consult the psychological literature on human causal inference, which attempts to model how people seem to perceive causality in events [65–67] and which produces empirical findings that are relevant to several of the common forms of elicitation observed in our analysis. For example, empirical research on the extent to which a model of causal support [66] captures idiosyncrasies of human causal inference suggests that people may find it more natural to distinguish which features have some causal effect than to estimate the strength of those effects. Such findings have direct implications for elicitation of relevant expert mental models for applied machine learning.

5.2.4. Establishing Context and Common Ground

Not all papers mentioned how exactly domain experts were introduced to the ML engineers' or researchers' goals, value judgments, or background perspectives on the models they were building. However, several recent qualitative studies suggest that these types of high level information about a model may be particular helpful to domain experts who will be end-users [38,39].

Graphical elicitation, in which visualization interfaces are used for elicitation, may be another fruitful way to help establish context as an expert provides their knowledge. Visualization interfaces that both capture new information from an expert, such as labels or interesting instances, and provide context, could help make elicitation more systematic. In other areas where elicitation is relevant, such as Bayesian models of cognition for graphical inference, graphical elicitation interfaces have been found to be useful for reducing abstraction in the elicitation process, such as when eliciting priors or beliefs about parameters [23,68,69]. Visualization interfaces might simultaneously be used to visualize back to an expert alternative representations of the elicited knowledge for validation.

Finally, related to providing context, some recent interactive machine learning research argues that, when experts are shown labeled data used for training or validation during elicitation, the knowledge that is elicited from them can be redundant in ways that hamper model performance [52]. This possibility is largely not mentioned by papers in our sample that explicitly described giving domain experts access to training data, but as elicitation interfaces become a more focal aspect of applied machine learning research, this risk may be important for researchers to account for.

5.2.5. Cognitive Bias

Of the 28 papers we analyzed in our survey, 5 of them mentioned the possibility of the experts exhibiting cognitive biases, and 4 of them explicitly attempted to mitigate bias when eliciting knowledge. The broader elicitation literature including prior elicitation and elicitation of beliefs or judgments in the judgment and decision making literature provides evidence of how researchers can obtain higher quality knowledge if they anticipate cognitive biases that the experts have and seek to minimize them during elicitation. For example, O'Hagan [5] notes that experts tend to be overconfident and are susceptible to the effects

of anchoring and availability heuristics. Of the four works in our survey that explicitly mitigated cognitive biases, two of them [13,18] presented labeling tools that anonymized labeler identities during discussions of disagreements. Yang et al. [6] conceptualized an interactive machine learning tool that first has the user select instances to use as test cases, guiding them away from a biased test set. Cano et al. [20] noted that eliciting expert knowledge about specific uncertain edges in a Bayesian network is preferable to having the expert provide their knowledge on edges up front because in that case the "expert could be biased towards providing the 'easiest' or clearest knowledge", which is likely already apparent in the data and easy for the model to learn. These examples are encouraging, but are exceptions among the papers that we analyzed.

5.2.6. Validation of Elicited Information

Only about one quarter of the paths (18/73) used some form of validation procedure on the elicited information to increase confidence in its correctness. Given that articulating one's knowledge is often challenging for experts even given well-designed elicitation processes, the relative lack of attempts to validate experts' knowledge (e.g., through presenting it back to them for discussion and confirmation, aggregating it with other experts' responses, ensuring that it was consistent upon repeated elicitation, etc.) presents an important area for improvement.

Where we did see validation, it often occurred through aggregation of responses from multiple experts. One relatively simple way for machine learning researchers to obtain more reliable and accurate knowledge from domain experts is by defaulting to multiple expert elicitation practices wherever possible. When multiple experts are used, there should be a procedure to aggregate that knowledge into a single answer. O'Hagan [5] explains that this can be achieved through behavioral aggregation by having the experts reach a consensus, which is the approach used in the SHELF protocol, or through mathematical aggregation that uses a pooling rule to combine expert answers, which is the approach used in the Cooke protocol. Even in situations where a probability distribution is not being elicited, there should be a planned way to integrate knowledge from multiple experts. Seymoens et al. [17] and Schaekermann et al. [18] provide two examples of behavioral aggregation, while Bowles et al. [13] and Ashdown et al. [9] give examples of using majority vote as a simple pooling rule. Seymoens et al. also addressed social dynamics that can arise during behavioral aggregation, as O'Hagan recommends, by checking if the participants felt that they were able to share their concerns.

Confirming elicited information between multiple experts is not the only way that the knowledge can be validated, however. O'Hagan recommends presenting the experts' answers to them in different ways in order to get them to think about the prompts from multiple angles, which allows the experts to conduct basic sanity checks on their answers [59]. For example, when eliciting the median value of a probability distribution, one way to encourage the experts to check their answer is to have them consider that, if they were to be given a prize if they guessed if the true value is above or below their stated median, then they should be indifferent to either option [5].

6. Future Work and Limitations

Our work is novel in its focus on how elicitation of domain expertise is practiced in applied ML settings. One obvious opportunity for future research is in systematizing and evaluating elicitation approaches. Our literature searches focused on elicitation in machine learning turned up little research attempting to establish how well a particular elicitation method worked. While it is difficult to judge the validity of elicited knowledge given that the target is typically based in the mental model of the expert, elicitation research outside of machine learning points to various possibilities, including endowing knowledge to test methods for eliciting it back, presenting simulations or implications of elicited knowledge to see how much an expert confirms when asked to consider it more deeply, and using repeated elicitation to confirm consistent responses from a method. There are also

opportunities to expand on past research in experts systems [46] by comparing, for example, the association of different methods with the impact of the work.

One interesting finding of our analysis is the extent to which the same form of information, like labels, or instances, or rules, could be elicited for very different purposes representing different points in the model pipeline. Given how little rich description of how context and common ground were provided, a researcher might easily see these elicitation attempts as more or less synonymous, assuming that prompts or interfaces for eliciting instances, for example, for one goal like problem specification can be applied equally well to other uses. How a domain expert understands the rationale for the elicitation, and how context on the modeling is provided, are however critical details that require sensitivity to the specific targets and pipeline if the elicited knowledge is to be reflective of the experts' actual mental model. One important area for future work is to better typologize different approaches to eliciting and using common formats like instances or feature relationships, distinguishing between what may otherwise be overlooked nuances in their use, since these appear often in the literature in applied ML as well as topics like ML interpretability, but can play very different roles in a modeling pipeline. Future research could also strengthen the connections between the empirical literature on human reasoning about causation and elicitation approaches to deepen motivation to elicit particular information over other information.

Future work could also survey additional papers in a similar vein to our analysis. Our sample is not a complete assessment of the elicitation practice in applied ML, due to the challenges of ensuring comprehensiveness given differences in the domains in which applied ML research is published and language used to describe elicitation, knowledge, and other concepts. As such, we encourage more observational research on how expert knowledge is being elicited and used for machine learning. There is a need for more research to identify and address the unique elicitation challenges for machine learning and to understand what lessons can be transferred from more traditional elicitation domains. This is particularly true for elicitation for deep learning, which we feel is underdeveloped compared to understanding elicitation for Bayesian or linear models. For example, 5 of the 28 papers that we coded mentioned using neural networks models, which is not wholly representative, given the diversity of model architectures, input types, and tasks that we see in modern deep learning. In addition, while we have included papers focusing on a variety of domains, such as medicine, energy systems, and water networks, there are likely many more instances of domain-specific applications of machine learning in domains that we did not consider.

A further limitation of our work is that it only analyzes explicit elicitation of domain knowledge. Our taxonomy does not cover implicit uses of domain knowledge, such as by domain experts that are a part of the research team and are tightly involved throughout the machine learning process. To better understand the dynamics of how expert knowledge is shared and used in these inter-disciplinary teams, more work along the lines of Mao et al.'s interview study [37] is needed.

7. Conclusions

Eliciting expert knowledge is a standard part of many machine learning workflows. We characterized the ways in which machine learning researchers elicit expert knowledge by developing a taxonomy that categorizes an elicitation approach according to the *elicitation goal*, *elicitation target*, *elicitation process*, and the *use of elicited knowledge*. Our survey coded 73 elicitation paths found across 28 papers and analyzed the trends that emerged in these paths when comparing the paths where elicitation was performed for *problem specification*, *feature engineering*, *model development*, and *model evaluation*. We identified gaps in these paths and motivated an increased focus on transparent, traceable, and systematic elicitation in applied machine learning.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/mti5120073/s1, Spreadsheet S1: Papers, Taxonomy, and Elicitation Paths, Software S1:

Visualizations. Spreadsheet S1: Papers, Taxonomy, and Elicitation Paths contains three sheets. The first sheet lists the 28 papers covered in our survey. The second sheet outlines the hierarchy of the taxonomy. The third sheet contains the 73 elicitation paths coded according to the taxonomy. Software S1: Visualizations contains the code used to create the visualizations in this paper. The visualizations were created using D3 [70] on https://observablehq.com (last accessed on 19 November 2021).

Author Contributions: Conceptualization, D.K., J.H. and E.B.; methodology, D.K., J.H. and E.B.; investigation, D.K.; writing—original draft preparation, D.K. and J.H.; writing—review and editing, D.K., J.H. and E.B.; visualization, D.K.; supervision, J.H. and E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the Supplementary Material.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Chi, M.T. Laboratory methods for assessing experts' and novices' knowledge. In *The Cambridge Handbook of Expertise and Expert Performance*; Cambridge University Press: Cambridge, UK, 2006; pp. 167–184.
2. O'Hagan, A.; Buck, C.E.; Daneshkhah, A.; Eiser, J.R.; Garthwaite, P.H.; Jenkinson, D.J.; Oakley, J.E.; Rakow, T. *Uncertain Judgements: Eliciting Experts' Probabilities*; John Wiley & Sons: Hoboken, NJ, USA, 2006.
3. Cartwright, M.; Seals, A.; Salamon, J.; Williams, A.; Mikloska, S.; MacConnell, D.; Law, E.; Bello, J.P.; Nov, O. Seeing Sound: Investigating the Effects of Visualizations and Complexity on Crowdsourced Audio Annotations. *Proc. ACM Hum.-Comput. Interact.* **2017**, *1*, 1–21. [CrossRef]
4. Cakmak, M.; Thomaz, A.L. Designing robot learners that ask good questions. In Proceedings of the 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Boston, MA, USA, 5–8 March 2012; pp. 17–24.
5. O'Hagan, A. Expert Knowledge Elicitation: Subjective but Scientific. *Am. Stat.* **2019**, *73*, 69–81. [CrossRef]
6. Yang, Q.; Suh, J.; Chen, N.C.; Ramos, G. Grounding Interactive Machine Learning Tool Design in How Non-Experts Actually Build Models. In Proceedings of the 2018 Designing Interactive Systems Conference, Hong Kong, China, 9–13 June 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 573–584. [CrossRef]
7. Sundin, I.; Peltola, T.; Micallef, L.; Afrabandpey, H.; Soare, M.; Mamun Majumder, M.; Daee, P.; He, C.; Serim, B.; Havulinna, A.; et al. Improving genomics-based predictions for precision medicine through active elicitation of expert knowledge. *Bioinformatics* **2018**, *34*, i395–i403. [CrossRef] [PubMed]
8. Madigan, D.; Gavrin, J.; Raftery, A.E. Eliciting prior information to enhance the predictive performance of bayesian graphical models. *Commun. Stat.-Theory Methods* **1995**, *24*, 2271–2292. [CrossRef]
9. Ashdown, G.W.; Dimon, M.; Fan, M.; Terán, F.S.R.; Witmer, K.; Gaboriau, D.C.A.; Armstrong, Z.; Ando, D.M.; Baum, J. A machine learning approach to define antimalarial drug action from heterogeneous cell-based screens. *Sci. Adv.* **2020**, *6*, eaba9338. [CrossRef]
10. Ustun, B.; Adler, L.A.; Rudin, C.; Faraone, S.V.; Spencer, T.J.; Berglund, P.; Gruber, M.J.; Kessler, R.C. The World Health Organization Adult Attention-Deficit/Hyperactivity Disorder Self-Report Screening Scale for DSM-5. *JAMA Psychiatry* **2017**, *74*, 520–526. [CrossRef]
11. Sendak, M.; Elish, M.C.; Gao, M.; Futoma, J.; Ratliff, W.; Nichols, M.; Bedoya, A.; Balu, S.; O'Brien, C. "The Human Body is a Black Box": Supporting Clinical Decision-Making with Deep Learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 99–109. [CrossRef]
12. Bowles, K.H.; Potashnik, S.; Ratcliffe, S.J.; Rosenberg, M.; Shih, N.W.; Topaz, M.; Holmes, J.H.; Naylor, M.D. Conducting research using the electronic health record across multi-hospital systems: Semantic harmonization implications for administrators. *J. Nurs. Adm.* **2013**, *43*, 355–360. [CrossRef]
13. Bowles, K.H.; Ratcliffe, S.; Potashnik, S.; Topaz, M.; Holmes, J.; Shih, N.W.; Naylor, M.D. Using Electronic Case Summaries to Elicit Multi-Disciplinary Expert Knowledge about Referrals to Post-Acute Care. *Appl. Clin. Inform.* **2016**, *7*, 368–379. [CrossRef]
14. Heckerman, D.; Geiger, D.; Chickering, D.M. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Mach. Learn.* **1995**, *20*, 197–243. [CrossRef]

15. Cai, C.J.; Reif, E.; Hegde, N.; Hipp, J.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viegas, F.; Corrado, G.S.; Stumpe, M.C.; et al. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1–14.
16. Lee, M.H.; Siewiorek, D.P.; Smailagic, A.; Bernardino, A.; Bermúdez i Badia, S. Interactive Hybrid Approach to Combine Machine and Human Intelligence for Personalized Rehabilitation Assessment. In Proceedings of the ACM Conference on Health, Inference, and Learning, Toronto, ON, Canada, 2–4 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 160–169. [CrossRef]
17. Seymoens, T.; Ongenae, F.; Jacobs, A.; Verstichel, S.; Ackaert, A. A Methodology to Involve Domain Experts and Machine Learning Techniques in the Design of Human-Centered Algorithms. In *Human Work Interaction Design. Designing Engaging Automation*; Barricelli, B.R., Roto, V., Clemmensen, T., Campos, P., Lopes, A., Gonçalves, F., Abdelnour-Nocera, J., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 200–214.
18. Schaekermann, M.; Hammel, N.; Terry, M.; Ali, T.K.; Liu, Y.; Basham, B.; Campana, B.; Chen, W.; Ji, X.; Krause, J.; et al. Remote Tool-Based Adjudication for Grading Diabetic Retinopathy. *Transl. Vis. Sci. Technol.* **2019**, *8*, 40. [CrossRef]
19. Masegosa, A.R.; Moral, S. An interactive approach for Bayesian network learning using domain/expert knowledge. *Int. J. Approx. Reason.* **2013**, *54*, 1168–1181. [CrossRef]
20. Cano, A.; Masegosa, A.R.; Moral, S. A Method for Integrating Expert Knowledge When Learning Bayesian Networks From Data. *IEEE Trans. Syst. Man Cybern. Part B* **2011**, *41*, 1382–1394. [CrossRef]
21. Richardson, M.; Domingos, P. Learning with Knowledge from Multiple Experts. In Proceedings of the Twentieth International Conference on Machine Learning, Washington, DC, USA, 21–24 August 2003; pp. 624–631.
22. Langseth, H.; Nielsen, T.D. Fusion of Domain Knowledge with Data for Structural Learning in Object Oriented Domains. *J. Mach. Learn. Res.* **2003**, *4*, 339–368.
23. Afrabandpey, H.; Peltola, T.; Kaski, S. Human-in-the-loop Active Covariance Learning for Improving Prediction in Small Data Sets. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, Macao, China, 10–16 August 2019; pp. 1959–1966. [CrossRef]
24. El-Assady, M.; Sperrle, F.; Deussen, O.; Keim, D.; Collins, C. Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution. *IEEE Trans. Vis. Comput. Graph.* **2019**, *25*, 374–384. [CrossRef]
25. El-Assady, M.; Kehlbeck, R.; Collins, C.; Keim, D.; Deussen, O. Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. *IEEE Trans. Vis. Comput. Graph.* **2020**, *26*, 1001–1011. [CrossRef]
26. Daee, P.; Peltola, T.; Soare, M.; Kaski, S. Knowledge elicitation via sequential probabilistic inference for high-dimensional prediction. *Mach. Learn.* **2017**, *106*, 1599–1620. [CrossRef]
27. Hu, R.; Granderson, J.; Auslander, D.; Agogino, A. Design of machine learning models with domain experts for automated sensor selection for energy fault detection. *Appl. Energy* **2019**, *235*, 117–128. [CrossRef]
28. Webb, G.I. Integrating machine learning with knowledge acquisition through direct interaction with domain experts. *Knowl.-Based Syst.* **1996**, *9*, 253–266. [CrossRef]
29. Camarinha-Matos, L.M.; Martinelli, F.J. Application of Machine Learning in Water Distribution Networks Assisted by Domain Experts. *J. Intell. Robot. Syst.* **1999**, *26*, 325–352. [CrossRef]
30. Ratner, A.; Bach, S.H.; Ehrenberg, H.; Fries, J.; Wu, S.; Ré, C. Snorkel: Rapid Training Data Creation with Weak Supervision. *Proc. VLDB Endow.* **2017**, *11*, 269–282. [CrossRef]
31. Ustun, B.; Rudin, C. Learning Optimized Risk Scores. *J. Mach. Learn. Res.* **2019**, *20*, 1–75.
32. Amershi, S.; Lee, B.; Kapoor, A.; Mahajan, R.; Christian, B. CueT: Human-Guided Fast and Accurate Network Alarm Triage. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Vancouver, BC, Canada, 7–12 May 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 157–166.
33. Altendorf, E.E.; Restificar, A.C.; Dietterich, T.G. Learning from Sparse Data by Exploiting Monotonicity Constraints. In Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence, Edinburgh, Scotland, 26–29 July 2005; AUAI Press: Arlington, Virginia, USA, 2005; pp. 18–26.
34. Holstein, K.; Wortman Vaughan, J.; Daumé, H., III; Dudik, M.; Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–16.
35. Kaur, H.; Nori, H.; Jenkins, S.; Caruana, R.; Wallach, H.; Wortman Vaughan, J. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25–30 April 2020; pp. 1–14.
36. Law, P.M.; Malik, S.; Du, F.; Sinha, M. Designing Tools for Semi-Automated Detection of Machine Learning Biases: An Interview Study. *arXiv* **2020**, arXiv:2003.07680.
37. Mao, Y.; Wang, D.; Muller, M.; Varshney, K.R.; Baldini, I.; Dugan, C.; Mojsilović, A. How Data ScientistsWork Together with Domain Experts in Scientific Collaborations: To Find the Right Answer or to Ask the Right Question? *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 1–23. [CrossRef]
38. Hong, S.R.; Hullman, J.; Bertini, E. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 1–26. [CrossRef]

39. Cai, C.J.; Winter, S.; Steiner, D.; Wilcox, L.; Terry, M. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 1–24. [CrossRef]
40. Ericsson, K.; Hoffman, R.; Kozbelt, A.; Williams, A. (Eds.) *The Cambridge Handbook of Expertise and Expert Performance*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2018. [CrossRef]
41. Chi, M.T.; Feltovich, P.J.; Glaser, R. Categorization and representation of physics problems by experts and novices. *Cogn. Sci.* **1981**, *5*, 121–152. [CrossRef]
42. Chi, M.T.; Glaser, R.; Rees, E. *Expertise in Problem Solving: Advances in the Psychology of Human Intelligence*; Erlbaum: Hillsdale, NJ, USA, 1982; pp. 1–75.
43. Garthwaite, P.H.; Kadane, J.B.; O'Hagan, A. Statistical methods for eliciting probability distributions. *J. Am. Stat. Assoc.* **2005**, *100*, 680–701. [CrossRef]
44. Goldstein, D.G.; Rothschild, D. Lay understanding of probability distributions. *Judgm. Decis. Mak.* **2014**, *9*, 1–14.
45. Wagner, W.P.; Najdawi, M.K.; Chung, Q. Selection of knowledge acquisition techniques based upon the problem domain characteristics of production and operations management expert systems. *Expert Syst.* **2001**, *18*, 76–87. [CrossRef]
46. Wagner, W.P. Trends in expert system development: A longitudinal content analysis of over thirty years of expert system case studies. *Expert Syst. Appl.* **2017**, *76*, 85–96. [CrossRef]
47. Rahman, P.; Nandi, A.; Hebert, C. Amplifying Domain Expertise in Clinical Data Pipelines. *JMIR Med. Inform.* **2020**, *8*, e19612. [CrossRef] [PubMed]
48. Amershi, S.; Cakmak, M.; Knox, W.B.; Kulesza, T. Power to the people: The role of humans in interactive machine learning. *AI Mag.* **2014**, *35*, 105–120. [CrossRef]
49. Cakmak, M.; Chao, C.; Thomaz, A.L. Designing interactions for robot active learners. *IEEE Trans. Auton. Ment. Dev.* **2010**, *2*, 108–118. [CrossRef]
50. Thomaz, A.L.; Breazeal, C. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artif. Intell.* **2008**, *172*, 716–737. [CrossRef]
51. Rosenthal, S.L.; Dey, A.K. Towards maximizing the accuracy of human-labeled sensor data. In Proceedings of the 15th International Conference on Intelligent User Interfaces, Hong Kong, China, 7–10 February 2010; pp. 259–268.
52. Daee, P.; Peltola, T.; Vehtari, A.; Kaski, S. User modelling for avoiding overfitting in interactive knowledge elicitation for prediction. In Proceedings of the 23rd International Conference on Intelligent User Interfaces, Tokyo, Japan, 7–11 March 2018; pp. 305–310.
53. Budd, S.; Robinson, E.C.; Kainz, B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med. Image Anal.* **2021**, *71*, 102062. [CrossRef]
54. Wu, X.; Xiao, L.; Sun, Y.; Zhang, J.; Ma, T.; He, L. A Survey of Human-in-the-loop for Machine Learning. *arXiv* **2021**, arXiv:2108.00941.
55. Lasecki, W.S.; Rzeszotarski, J.M.; Marcus, A.; Bigham, J.P. The Effects of Sequence and Delay on Crowd Work. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, Seoul, Korea, 18–23 April 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1375–1378. [CrossRef]
56. Attenberg, J.; Ipeirotis, P.G.; Provost, F.J. Beat the Machine: Challenging Workers to Find the Unknown Unknowns. In Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11), San Francisco, CA, USA, 7–8 August 2011; Volume WS-11-11.
57. Lofland, J.; Lofland, L.H. *Analyzing Social Settings*; Wadsworth Pub. Co.: Belmont, CA, USA, 1971.
58. Clark, H.H.; Schreuder, R.; Buttrick, S. Common ground at the understanding of demonstrative reference. *J. Verbal Learn. Verbal Behav.* **1983**, *22*, 245–258. [CrossRef]
59. O'Hagan, A. Probabilistic Uncertainty Specification: Overview, Elaboration Techniques and Their Application to a Mechanistic Model of Carbon Flux. *Environ. Model. Softw.* **2012**, *36*, 35–48. [CrossRef]
60. O'Hagan, A.; Oakley, J.E. *SHELF: The Sheffield Elicitation Framework (Version 4)*; University of Sheffield: Sheffield, UK, 2019.
61. Gosling, J.P. SHELF: The Sheffield Elicitation Framework. In *Elicitation: The Science and Art of Structuring Judgement*; Dias, L.C., Morton, A., Quigley, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 61–93. [CrossRef]
62. Cooke, R.M. *Experts in Uncertainty: Opinion and Subjective Probability in Science*; Oxford University Press: Oxford, UK, 1991; p. xii-321.
63. Rowe, G.; Wright, G. The Delphi technique as a forecasting tool: Issues and analysis. *Int. J. Forecast.* **1999**, *15*, 353–375. [CrossRef]
64. Hullman, J.; Kay, M.; Kim, Y.S.; Shrestha, S. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE Trans. Vis. Comput. Graph.* **2017**, *24*, 446–456.
65. Cheng, P.W. From covariation to causation: A causal power theory. *Psychol. Rev.* **1997**, *104*, 367. [CrossRef]
66. Griffiths, T.L.; Tenenbaum, J.B. Structure and strength in causal induction. *Cogn. Psychol.* **2005**, *51*, 334–384. [CrossRef]
67. Griffiths, T.L.; Tenenbaum, J.B. Theory-based causal induction. *Psychol. Rev.* **2009**, *116*, 661. [CrossRef]
68. Hullman, J.; Gelman, A. Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference. *Harvard Data Science Review* **2021**. [CrossRef]
69. Kim, Y.S.; Walls, L.A.; Krafft, P.; Hullman, J. A bayesian cognition approach to improve data visualization. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, UK, 4–9 May 2019; pp. 1–14.
70. Bostock, M.; Ogievetsky, V.; Heer, J. D^3 Data-Driven Documents. *IEEE Trans. Vis. Comput. Graph.* **2011**, *17*, 2301–2309. [CrossRef]