The Genie in the Bottle: Different Stakeholders, Different Interpretations of Machine Learning

Mahboobeh Harandi
Syracuse University
mharandi@syr.edu

Kevin Crowston
Syracuse University
crowston@syr.edu

Corey Jackson
University of California, Berkeley
coreyjackson@berkeley.edu

Carsten Østerlund
Syracuse University
costerlu@syr.edu

Abstract

We explore how people developing or using a system with a machine-learning (ML) component come to understand the capabilities and challenges of ML. We draw on the social construction of technology (SCOT) tradition to frame our analysis of interviews and discussion board posts involving designers and users of a ML-supported citizen-science crowdsourcing project named Gravity Spy. We extend SCOT by anchoring our investigation in the different uses of the technology. We find that the type of understandings achieved by groups having less interaction with the technology is shaped more by outside influences and less by the specifics of the system and its role in the project. This initial understanding of how different participants understand and engage with ML points to challenges that need to be overcome to help users of a system deal with the opaque position that ML often holds in a work system.

1. Introduction

Machine learning (ML) has recently increased in capability and is being more broadly applied. However, the application of ML has distinctive characteristics that are unlike other approaches for supporting or automating work. First, most ML systems are trained using pattern recognition techniques rather than explicitly programmed. For instance, the performance of supervised learning algorithms is heavily dependent on the quantity and quality of data available for training the model and classification. Second, the results of ML are most often probabilistic, that is, when classifying an unknown case, the ML output provides probabilities that the unknown case fits one or some of the known categories rather than a definitive answer. Finally, the processes of some ML techniques (e.g., neural networks) are opaque, that is, it is challengingto explain why a particular output was selected.

The differences described above may cause problems for use and users of the technology. The application of an ML system is clearly an algorithmic phenomenon, but human ability to control the technology is limited. For instance, an unwanted behavior is harder to fix if it is the result of a biased training dataset rather than the algorithm design itself.

Given the challenges associated with interpreting ML, we are interested in how people, particularly non-experts, make sense of ML in work systems. The question we explore in this paper is:

> how do people developing or using an ML system realize the distinctive characteristics and limitations of the technology?

We explore this question in the context of an online citizen-science project called Gravity Spy (`http://www.gravityspy.org`) that incorporates ML and involves a number of different users groups having varied interactions with the technology, thus providing a diversity of perspectives on ML.

2. Theory

We build our exploration on how people approach, work with and perceive ML on two basic concepts, interpretive flexibility and relevant social groups, as discussed in the Social Construction of Technology (SCOT) program [1]. Emerging out of the "Bath school" in Science and Technology Studies by Collins [2] and his students, Pinch and Travis, this human-centered approach is concerned with the design actions taken by different groups and the meanings these impart. SCOT has had some influence on the general proliferation of science and technology studies across the social sciences and in information systems more specifically. Many papers offer broader theoretical introductions to the framework (e.g., [3, 4]), but we also find empirical studies drawing on the framework (e.g., [5, 6]).

First, the notion of interpretive flexibility highlights that technologies and artifacts can be associated with more than one meaning. Sufficiently underdetermined, technological artifacts allow for multiple interpretations

and possible designs. Concerned with the social shaping of science and technology, Collins [2] and later Pinch and Bijker [1] suggested that technology design is an open process with different potential outcomes depending on the social circumstances of its development.

Second, the concept of relevant social group embodies people with a common interpretation, that is, all members of a certain social group that share the same set of meanings attached to a specific artifact [1, p. 414]. To determine who falls into such a group, Pinch and Bijker [1] ask a series of questions.

- First, does the artifact have any meaning to the members of the social group under investigation? Obvious groups include users or consumers of an artifact but there can be less obvious groups.
- Second, is a previously-defined social group homogeneous when it comes to the meanings given to an artifact or would it be helpful to break a heterogeneous group into sub-groups?
- Third, in defining relevant social groups, Pinch and Bijker [1] are focused on the problems facing each group in relation to the artifact.
- Finally, a number of technological solutions might emerge around each problem.

By focusing on problems and solutions, Pinch and Bijker [1] do not go into details about the type of practices associated with the artifact and how groups may engage with an artifact in radically different ways, though this perspectives seems to be behind the perception of problems.

Finally, in Pinch and Bijker [1], interpretive flexibility is eventually overtaken by processes of closure and stabilization. However, as we are studying a technology as it is newly deployed, we do not expect to see this part of the process.

In summary, the SCOT approach to technology suggests identifying the relevant social groups around a technology by looking for groups with relatively homogeneous perceptions of the problems with a technology and the solutions for those problems. We extend this approach by first considering how the groups may differ in how they interact with the technology that lead to perception of problems, as well as the resources they can draw on to develop their understandings of solutions.

3. The Gravity Spy System

We examine perceptions and the origins of users' perceptions of ML technology in the context of a citizen science project called Gravity Spy. Citizen science describes an arrangement where members of the public contribute to scientific research [7]. Gravity Spy supports research in the Laser Interferometer Gravitational-Wave Observatory (LIGO) scientific collaboration, a consortium of researchers and institutions working to record evidence of gravitational waves [8, 9]. The consortium uses detectors that, in addition to potential gravitational waves, record internal and external noise (called "glitches") produced as a result of the sensitivity level that is required to record gravitational waves. Since there are hundreds or thousands of glitches every day, human eyes are needed to classify glitches so they can be studied and their causes determined. Scientists ask volunteers on the Gravity Spy platform to classify glitches and in cases where no existing category exists, to propose a new category.

3.1. Gravity spy as a hybrid human-machine system

Gravity Spy incorporates ML in three ways.

- The ML Classifier: a supervised deep-learning classifier that was trained on gold data (i.e., glitches classified by experts). The ML classifies glitches as they are added to the system into one of the twenty-two known glitch classes. It provides the likelihood of the glitch belonging to each of the classes. The classifications are used to route glitches to volunteers, with beginners getting glitches for which the ML is more confident and more advanced users glitches with lower confidence that are presumably harder to classify.
- The search tool: ML is applied to support the process of finding new glitch classes. A similarity-search tools allow volunteers to search for glitches similar to a chosen glitch.
- Clustering: an unsupervised learning algorithm that identifies and groups similar "none of the above" glitches to propose new glitch classes.

The result is a hybrid human-machine system, using ML techniques intertwined with the dataset that has been used to make a predictive model for labeling unseen data. Since the training dataset for Gravity Spy was created by the science team, their interpretations and biases affect the process of agreement and quality of the training dataset. The quality of the training data in turn affects the process of feature selection, feature extraction and the ML algorithm's predictions. Further, each group of people have different interpretations of

how the ML algorithm has classified unseen data, as it is challenging to understand why it has predicted a specific result. Each group faces different problems depending on their interpretations and interactions and seeks different solutions to address the problems.

4. Research Methodology

A qualitative approach was adopted to understand how individuals approach, work with and perceive ML. Qualitative approaches have proven valuable in understanding the social and cultural significance people impart on technologies [10, 11, 12].

4.1. Data elicitation

The empirical data for our study come from two sources: interviews and the Gravity Spy discussion forum posts. We conducted six **interviews**: two with volunteers who are also moderators, and four with members of the Gravity Spy science team. The selection of interviewees was based on a purposive and opportunistic sampling procedure. From the volunteer population, we chose to interview volunteers who had been a part of the project for a long time allowing them to have come into contact with most ML components.

The goal of the interviews was to understand how interviewees understand and interact with ML in Gravity Spy. Using a semi-structured interview protocol we asked questions such as "Can you describe the functioning of the ML in the project and what role it plays in various stages of the work process?" Each interview was audio recorded and lasted approximately one hour and was transcribed.

As for the second source, **discussions** in Gravity Spy are diverse and cover a variety of topics written by volunteers and the Gravity Spy team. We collected comments (N = 249) posted by thirty-three volunteers to the discussion fora pertaining to the ML functions, use, problems and solutions. To find relevant posts, we conducted a keyword search on the Gravity Spy homepage. We broadened our search to include related terms such as: algorithm, machine learning, pattern recognition, machine teaching, computer learning, artificial intelligence, bot, chatbot, AI and ML.

Volunteers of all abilities as well as members of the Gravity Spy science team post comments to the discussion boards and thus offer insights from a broader group of participants spanning from newcomers to experienced volunteers. We selected the most relevant and representative comments to our research question posted by seven volunteers. Names (pseudonyms) of interviewees, participants in discussion fora, and their roles in the project are displayed in Table 1.

Table 1. Interview subjects, participants from discussion fora and their role in the project. Pseudonym were used to protect the identity of subjects

Name	Role in Project	Data Collection
Peter	Gravity Spy Team	Interview
Marsha	Gravity Spy Team	Interview
Casper	Gravity Spy Team	Interview
April	Gravity Spy Team	Interview
Brandon	Moderator	Interview
Katie	Moderator	Interview
Jacob	Moderator	Discussion Posts
Olivia	Volunteer	Discussion Posts
Emilia	Volunteer	Discussion Posts
Aryan	Volunteer	Discussion Posts
Cruz	Volunteer	Discussion Posts
Ava	Volunteer	Discussion Posts
Sophia	Volunteer	Discussion Posts

4.2. Data analysis

The data were analyzed using thematic analysis [13, 14] with SCOT as a sensitizing device. We started with SCOT concepts of relevant social groups and interpretive flexibility. Once the interviews were completed, the authors read through the interview notes and transcripts identifying patterns of use, work with, and perceptions of ML. Interviewee statements were captured and organized based on similarities in how they work with the ML. We reviewed each category and developed themes around how individuals used the ML and the problems they experienced. Individuals with common problems were then linked to the relevant social group. The themes that describe how they use the technology, what problems they have and how they solve problems are described in the results. We used the discussion posts to corroborate volunteer accounts.

5. Results

5.1. Roles of Stakeholders in Gravity Spy

We identified different social groups around each technology by considering how each group use these technologies in their work, what problems they face and how solve related issues. In doing so, we followed the SCOT strategy of starting with pre-existing social groups (such as Gravity Spy team) and breaking them up if they had different uses or perceived various problems or solutions or merge them into one group if they had similar uses, challenges, and solutions. We explain the roles of each stakeholder's groups as follows:

ML developers designed and developed the ML

algorithms with the aim of (1) classifying images to known classes and (2) find similar image for the search interface. Moreover, they have been working on designing ML algorithms to find new classes of data. Peter is identified in this group.

Platform developers build and maintain the Gravity Spy system hosted on the Zooniverse.org platform. One of the main roles of this group is converting raw glitch data provided by LIGO scientists to images that are perceivable by volunteers, ML algorithms, and the science team. In addition, they collaborate with the LIGO scientists and ML developers to integrate the gold data and the ML algorithms into the system. Casper and Marsh are identified in this group.

LIGO scientists benefit from the glitch data analyzed by volunteers in Gravity Spy. They have provided the learning material for volunteers about different glitch types. They have been collaborating with the ML and platform developers to improve the gold dataset that they have created in the beginning of the project. April is identified in this group.

Volunteers can be any individual with internet access who chooses to spend time analyzing glitch data and learning about the science behind the Gravity Spy project. They go through a scaffolded workflow that progressively presents them with more challenging classification tasks. Moderators and other expert volunteers alsoparticipate in discussions. Brandon, Katie, Jacob, Olivia, Emilia, Aryan, Cruz, Ava and Sophia are identified in this group.

5.2. Use of the ML Technologies

We examine three ML technologies in this paper, the ML classifier using labeled data as a supervised algorithm, the search tool, and the clustering algorithm as an unsupervised algorithm. These uses have different exposures to different users, making them interesting to compare. The classifier is embedded in the system in a way that is not visible to end users, while the search tool is an end-user tool. The clustering tool has not yet been deployed.

The ML Classifier

The first group, **ML developers** explained that they trained the ML classifier based on the gold data that was created by the LIGO scientists. Peter emphasized the importance of gold dataset as:

> We train the ML classifier based on the labeled data in golden set. ML cannot do a magic. ... Golden set is the heart of ML algorithms [in Gravity Spy]. We use those data to develop our ML algorithms for the Gravity Spy. The algorithms are in two main groups; the supervised group to classify data and score them that ... and unsupervised learning ...

The developers believe that the ML classifier works well for most of the known classes. However, they need volunteers to check the results of ML classifier.

Platform developers, the second group, first created a grounded format for data that is understandable by all parties, including volunteers, LIGO scientists, the science team, and ML developers. Casper said:

> I think the best thing we did with the machine learning from the volunteers' perspective and the LIGO perspective, is we presented the output in a really nicely digestible way, as images. People can just understand that better than the raw data. And ML also has an output in a grounded way, images.

Second, in collaboration with the LIGO scientists (see below), they classified glitches to create the initial version of the gold dataset. Third, they collaborated with the ML developers to integrate the ML classifier results into the platform. They designed the platform to assign images to different workflow based on its confidence score. However, for retiring images (that is, to assign a label for use in further studies), they consider that the volunteers' classifications are reliable and so see a need for volunteers to check the result of the ML classifier.

LIGO scientists, the third group, have been collaborating with the ML and platform developers and know how improving the gold dataset and the ML classifier had a positive impact on the results of the classifications. However, they otherwise do not know the details of the ML classifier. These scientists asked for and were given access to the results of the ML classifier while the system was in development, before the results of volunteers' classifications were available. As well, particular scientists have asked the platform developers to run the ML classifier on a set of images on a specific date to check quickly if they can find a correlation between the data and detector instruments. To explain this process, April said:

> There was something wrong in data and we asked Casper to run ML on yesterday's data and see how many whistles were there and what frequency and time they happened. We were able to do statistics on data and see hundreds of whistles happen at this time

> and we could look at the instruments at that time and see if there is any correlation.

The final group identified are the **volunteers**, who are affected by ML as its classifications govern what glitches they see. Some of them know what they classify in each workflow has already been classified by the ML classifier. Also, they know that the ML classifier is supposed to learn from volunteers by aggregating their classifications for known classes. However, there are some volunteers who want to know if their contributions really improve the ML classifier. Aryan posted on the discussion board:

> Do you have some insight into the effect or lack of effect the human classifications are having on the ongoing machine learning? I'd really like to see more feedback from the LIGO team to help me justify spending my time in this endeavor.

A few of them think that ML classifier should learn the way that they are doing the classification. Cruz left a comment on the discussion board:

> I am struggling with this idea a bit... I think that the ML is the one who should learn to adapt to us and not vice versa.

And a few of them tried to learn how the ML classifier would classify an image to make sure that they are classifying correctly and if the ML classifier has something to teach them.

The Search Tool

The **ML developers** also created the algorithms for the search tool. The feature extraction in the ML classifier was used as the basis for an algorithm to find similar images to a target provided by a user. Besides, they used the gold dataset to train the algorithm to find similar images in potential new classes.

The **platform developers** designed the search system and an interface to the tool, which enables volunteers to expand their collections of images. They believe it is a complementary tool for volunteers to accelerate the process of finding new classes.

The **LIGO scientists** know of the tool but do not the use it themselves. They think it should speed up the process of finding new classes for volunteers.

A few of the **volunteers** have been testing the search tool to expand their collections of known and unknown classes. They use it differently based on their perspectives of the ML classifier. One tested the search tool and reported bugs to the platform developers; another used it to search for images for which he thinks the ML classifier has a high confidence score.

Clustering

The **ML developers** designed and developed clustering algorithms to find new classes, experimenting with different approaches to the problem of unsupervised clustering of the images. They have applied two different techniques to develop the clustering tool. The former is relying on gold dataset to find new classes. They transferred gold dataset to the framework of clustering to find new classes. The second technique is using expert volunteers' collections as initial seeds for clustering to find clusters of new classes. However, these algorithm have not yet shown adequate performance and have not been deployed.

The **platform developers** are collaborating with ML developers to integrate the result of clustering into the search engine. They would like to know if the clustering is able to find new classes. They said:

> I think [the ML developers] managed to improve the model for clustering, even though it was computationally expensive, and I'm going to run it and just see if that helps with some of these searches for new classes.

In order to deploy the clustering, the platform developers will need to develop an interface to make the results available to volunteers.

The **science team** and the **LIGO scientists** know that the ML developers are developing the tool to find new glitch classes but they are not using clustering in their work. Rather, they currently rely on the volunteers' findings to discover new classes. They have checked different proposals submitted by volunteers very precisely to see if they agree with what they have proposed as a new class. So far, they have approved a few new classes that have been proposed by volunteers and added them to the list of known classes in Gravity Spy project.

Volunteers are at the present not informed about the possibility of clustering as it has yet to be deployed or even beta tested. They are the ones who have finding new classes but so far not the clustering. As they are going through different images in upper workflows they would propose new classes to the LIGO scientists but they do not know that the clustering has been developed to find new classes.

5.2.1. Other perceived uses of the technology In addition to the three implemented functions described above, some of the **volunteers** believed that an ML system has been used to communicate with volunteers

on the talk page. Brandon and Katie are identified in this group. Brandon said:

> I read on the Oxford websites ... that they plan in the future to teach autonomous agents who can talk different projects on Zooniverse. Even in Zooniverse, on the main talk forums there are talks about this. So, I think it's happening.

He added that one of the users on Gravity Spy commented on an image in a way that indicates it is a bot. He thinks humans would analyze the image in a different way than what was said about the image.

5.3. Problems with the technology

In this section we describe the problems with the technology as perceived by the members of each of the identified social groups.

The ML Classifier

ML developers faced different problems designing and implementing the ML classifier. They said that in the early phases of the project they had to retrain the ML model several times on new versions of the gold dataset given by the LIGO scientists, which was computationally expensive. They explained since the ML classifier relies on the gold dataset to learn the classification, it is necessary to retrain it when there is a new gold dataset, but that doing so takes quite a lot computational resources. Later they faced misclassification of some images by the ML classifier caused by some erroneous labels in the gold dataset and an error of the ML classifier.

A first problem of the **platform developers** was to present the results of ML classifier to volunteers or LIGO scientists in an understandable way. Later, they also noticed the misclassification problem of the ML classifier. They said the problem in gold dataset and the ML classifier's algorithm caused the misclassification. But there are also images that could fall in several categories that cause the misclassification. And the current issue is to design a right schema for weighting volunteers' classifications and integrate that to the system to retrain the ML classifier. Making decision on including what parameters is challenging. It affects the score of ML classifier and they need to come up with a framework that has the best impact on the score of the ML classifier. Marsha stated:

> I think that's probably the biggest challenge from the people side is how we can adequately cover all the different parameters that we're able to change and get an understanding of what really affects the results the most.

LIGO scientists were also aware of the primary misclassification of data by the ML classifier and knew the reason was the gold dataset and the ML classifier. They do not have any issues with the current ML classifier and indeed, already use its outputs.

Volunteers also knew about the wrong data in gold dataset that was causing the misclassification by the ML classifier. And there are some volunteers who are concerned if the problem of the ML classifier decreased by training on newcomers' classifications. Olivia posted a comment on the discussion board:

> I'd be surprised to learn that GS' problems likely really messed with at least some newbie's classifications. Did those messed up classifications, in turn, mess with the way the ML worked during that time?

She also believes that the ML classifier is not working well for all classes especially in upper level where it does not have a high confidence score and images can fall into several classes.

The Search Tool

ML and platform developers have faced the same problem in the search tool: it does not retrieve relevant images when searching for an image that does not belong to the known classes. Instead, the result is either nothing or a non-matching image.

Only a few of the **volunteers** have worked with the search tool and they found it very challenging, as the way the ML classifier sees glitches is quite different from how humans see them, leading to images being retrieved that do not seem similar to the volunteers. Brandon said:

> if something is completely new and unknown for the machine learning, it will not be, necessarily, clustered together, but it's around in the neighboring clusters, probably, or even not in the neighboring clusters but in different clusters having common features with the one I was looking for.

Another volunteer could not use the search tool because she could not get satisfying results for images that do not belong to the known classes. Katie said:

> I'm eagerly waiting to see how it is developed so I can use it. I'm not able to

> use it yet in a reliable manner, but I did get notified by Casper, letting me know that he's working on it.

Clustering

ML developers identified problems for designing the clustering to discover new glitch classes. As there are no instances for unknown classes in a training dataset, the clustering should learn to find new clusters without having any training (gold) dataset. It is very challenging to get right clusters of unknown classes. They said:

> The performance of ML for discovery of new glitch classes is not as good as ML classifiers for known classes. ML classifiers are trained for known classes but there is no ground truth for ML to discover new classes.

The **platform developers** have other issues on discovery of new glitch classes by the clustering. They needed to create a new infrastructure for discovery of new classes. They said:

> On the other hand, though, like I was saying, the novel classes of glitches are something that really we've only built the infrastructure for the volunteers to explore and we're still working to build the infrastructure for the machine to explore, too.

Besides, they need to create a user-friendly interface to present what the clustering has identified as a potential new class so volunteers can evaluate the result. Regarding improvements to the clustering results, they said there are some new classes that may have some images but it is not computationally efficient to train algorithm for few samples.

The **science team** know that the clustering cannot yet find new classes and they should have a new infrastructure for that. They know that the clustering needs the volunteers' help to find new classes. They know the current tools on the Gravity Spy are not supporting volunteers to find new glitch classes and they needed to design a complementary tool to support their works on finding new classes.

LIGO scientists rely on what volunteers proposed as a new glitch class and do not have any results from the clustering. However, as they are satisfied with the ML classifier, they are optimistic about the clustering and think it should work well soon.

Volunteers think that the ML cannot find new classes. Some of them strongly believe they should define more fine-grained classes that will provide the clustering algorithm with better data with which to find new classes.

5.4. Solutions to problems with the technology

Finally, we discuss what members of each of groups perceived as potential solutions to the identified problems.

The ML Classifier

The **ML developers** improved the algorithm of the ML classifier to handle the problem of the misclassification and trained it over the new gold dataset with corrected labels. This work resolved the problem of misclassification. They know it is expected that the ML classifier classifies all images to the right classes and they used state-of-the-art algorithms to make it more accurate and sufficient. The ML developers think that the ML classifier would have a different result if they add to the training data the glitches that the volunteers have classified. However, they have not yet retrained the ML classifier with the volunteers' data. They believe it is ambitious to not evaluate the results of the ML classifier and trust it without volunteers' evaluations.

The **platform developers** also believe it would be ideal to have the perfect algorithms for ML classifier that are able to classify the images without any needs for evaluations. However, they tried to come up with some solutions to improve the ML classifier in GS. They created a framework to include all volunteers' classifications based on their expertise and assign a credit to each volunteer. They should work on that to see if it improves the result of the ML classifier. Casper said:

> I think my big interaction with machine learning has gone from a worry about it being too much of a black box that would still opaque our understanding of the data. Our understanding of the data was already opaque because there's too much of it, but I wasn't necessarily sure machine learning was going to solve that for us. But I think the machine learning with the whole nuance that we've given it through the project has, in fact, had that result.

The **LIGO scientists** helped to correct the gold dataset labels, which consequently improve the ML classifier. Since then they are very satisfied with the current results of the ML classifier.

Some **volunteers** approached their problems by understanding how the ML classifier can be improved

over known classes. They think they are interacting directly with the ML classifier's result in each workflow and learn what ML is classifying. Brandon said:

> ...I would have classified it in a different category. But I have accepted that the machine classified it that way, and during the learning process, I was trying to learn how the machine thinks. Because sometimes I could be wrong, too; other times, the machine could be wrong. And in each cases, it's especially unclear who is right. Sometimes you just decide.

Regarding images that fall in several classes he said he need a consensus for lots of cases as they should make a decision to have a ground truth for those images. There are other volunteers who try to understand how to improve the ML classifier by proposing some solutions. Emilia posted on the discussion board:

> Maybe the machine algorithms could have variables that are a function of weather or time of day or local temperature or magnetic field or whatever may affect the measurement.

Ava thinks an efficient pattern recognition algorithm can classify lots of different images correctly. She commented:

> It's entirely probable many of the different patterns are related by cause. The whole effort could be done by a sophisticated pattern recognition program.

There are volunteers who know that it takes a long time to have a perfect ML classifier and there should be huge amount of labeled data. Jacob said:

> The very nature of what we are doing here means the pre-sorting algorithm isn't going to be perfect until the project is over.

Sophia posted:

> We are training the system, and that is something that isn't thought about by the AI experts. Google's search engine has been trained by the end users billions of times a day. The gravity wave program does not have the trainers to reach that size.

The Search Tool

The **ML developers** retrained the ML classifier with a new feature and they believe it should improve the result of the search tool. However, Peter said they need to integrate expert volunteers collections into the algorithms to improve the search result for images that do not belong to the current classes.

The **platform developers** are positive to see how improving the ML classifier improve the search result. Casper said:

> I think Peter managed to improve the model for clustering, even though it was computationally expensive, and I'm going to run it and just see if that helps with some of these searches ... .

They think that the search tool will be a good solution to finding new classes. Consequently, volunteers' collections will be used to train the clustering to find new classes after the evaluations by volunteers. Marsha said:

> It's giving us a complementary method to help uncover these different classes. I think that the best case is going to be when the machine learning can pretty much immediately identify clusters of new classes and then the volunteers can go in and verify these clusters and verify these new types of glitch classes.

The **science team** hope the current search engine accelerate the process of finding new classes for volunteers that consequently helps ML developers to improve the clustering.

One of the **volunteers** explained how they handle the issue of finding similar images if they belong to a new class. They try to find images through what the algorithm already knows, Brandon said:

> But if something is very new, with new features, it doesn't recognize the new feature, but relates each to already known types, So, if I try to find similar glitches to something that I think is very new, I have to try to do some indirect searches, so I try to think a little like a machine: Okay, what kind of known features can be recognized on this new glitch?

Clustering

ML developers believe it is very ambitious to expect clustering to find new classes by itself without any further evaluations. To deal with their current problems for discovery of new classes, they used the current gold dataset to train ML to find new classes. Besides,

they plan to use the collections of expert volunteers for finding new classes. They said:

> The ML cannot do magic. We need labeled data for ML but since we did not have any labeled data for discovery new classes we transferred the gold dataset to the ML algorithms to find new classes. Besides, we will use collections of expert volunteers for this purpose.

The **platform developers** believe it would be ambitious to find new classes through the clustering. But aligned with this goal they developed a tool to help volunteers to search for similar images to what they have in their collections. They said:

> I'm really excited for this whole new Gravity Spy, using Gravity Spy tools to create these vetting workflows to help facilitate new models, new classes. I think it's going to really jazz the users, or I hope it does at least.

The **LIGO scientist** know that clustering should be able to find new classes but they do not know how much it is relying on collections of expert volunteers. They think advancement in the algorithm of clustering should be enough to find new classes.

Some **volunteers** believe that LIGO science team should define more fine grained classes that would help clustering to find new classes. "if we were really interested in training the machine, we would have many more categories." Another solution that they suggested to the ML developers and the science team is adding independent variables beside the feature of images to train ML to find new classes. Although some think it should have a more sophisticated pattern recognition algorithm, there are some volunteers who believe clustering cannot find new classes unless there are labeled data for the training.

6. Discussion

The case presented above has some implications for building systems that embody ML and for researching them. First, methodologically, we found that it was useful when documenting the relevant social groups and their perceived problems and solutions to consider what use members of the groups were trying to make of the technology and so their opportunities to learn about it. In the Gravity Spy case, the groups and their relation to the technology are shown graphically in Figure 1. The figure shows that the ML developers are closest to the new technology (the "genie in the bottle"), as

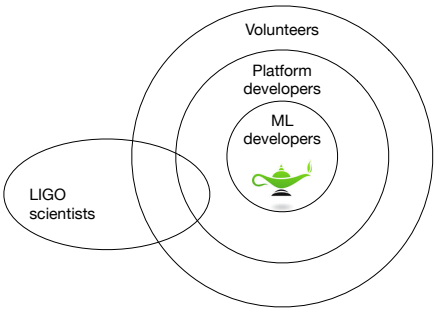

Figure 1. Circles of engagement with machine learning in Gravity Spy.

they are intimately involved in and try to make it work (coaxing the genie out of the bottle). However, other groups interact with the technology more indirectly. The volunteers, for example, are subject to the decisions of the ML classifier but have no easy way to see how it is designed or how it is performing. As a result, the further away from the bottle, the fuzzier the conception becomes.

Second, and related to the first point, groups with less contact with the technology must rely on other sources of information to make sense of its capabilities. For example, the LIGO scientists do not have the experience of building the ML classifier themselves, as Casper said:

> A lot of people just receive the end-products. There are some inputs. There's a black box and then there's some end-products and they don't think about either the inputs or the black box that led to the end-products. They look all GPS times have labels and people think it's ok. So, taking the output without knowing the input or the black box makes everything blurry.

In short, they see the output, classified glitches, which address a pressing need within their own practice, and not the caveats about performance.

Volunteers have even less opportunity to see how ML is being used as the system does not expose the details of the ML performances to avoid biasing volunteers' own classification. However, this design means that users have no easy way to explore the system's capabilities. Rather, it appears that in making sense of an "ML assistant", they draw on their own experience as contributors to the project, to scraps of information on various project blogs and to more general publications about AI.

A particular confusion seems to be about the difference between narrow and broad AI, i.e., a system

able to do just one task vs. one that can do many. This confusion leads some to conceive of the ML as filling the role of a participant in the project (i.e., anthropomorphism), not only classifying but also posting and discussing. As a result of this belief, there are interactions in which volunteers believe humans actions are actually those of machines (i.e., bots), what we label "technopomorphism". Given the rapidly advancing capabilities of chatbots, belief in chatbots is not unreasonable, and indeed, there may soon be Zooniverse chatbots, even though there are not at present. This experience suggests that when the bots do arrive, the identity of the human and machine elements should be made clearly visible to volunteers with labels in spaces where the two interact and tutorials describing where the boundaries of human and machine are, i.e., providing resources for understanding the genie even when it is not directly visible.

7. Conclusion

This initial study has examined just one setting with a limited number of interviews. In future work, we hope to expand to more settings and more thorough data collection. As well, our initial findings provide the basis for development of a systematic coding system for the volunteers' posts. Even in its initial state, we believe our study is useful in revealing the difficulties stakeholders in an ML may face in forming an accurate understanding of the system's role and capabilities. Misapprehensions about technology capability are not restricted to Gravity Spy. For example, witnessed by recent crashes, Tesla drivers seem not to universally understand the limits of the Tesla Autopilot (a problem that is not helped by choice of name). These understanding matter because the level of performance that is required or suitable depend heavily on the context. Some error in targeting an ad is okay, in diagnosing a disease less so and in recommending a prison sentence or driving a car, perhaps not at all. But from the outside, a user may not be able to tell how well a system for these different uses is performing. And conversely, the requirements that are apparent to users are less visible to developers, leading to a mismatch between design and expected performance. Future work should consider how to make the limitations of ML more visible to those who interact with its results but not the technology itself. It will be beneficial to have a standardized and easy to understand a way to communicate an ML system's level of performance, something akin to the descriptions of gas mileage found on cars.

References

[1] T. J. Pinch and W. E. Bijker, "The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other," *Social Studies of Science*, vol. 14, no. 3, pp. 399–441, 1984.

[2] H. M. Collins, "The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics," *Sociology*, vol. 9, no. 2, pp. 205–224, 1975.

[3] H. K. Klein and D. L. Kleinman, "The social construction of technology: Structural considerations," *Science, Technology, & Human Values*, vol. 27, no. 1, pp. 28–52, 2002.

[4] D. Howcroft, N. Mitev, and M. Wilson, "What we may learn from the social shaping of technology approach," *Social Theory and Philosophy for Information Systems*, pp. 329–371, 2004.

[5] S. Cadili and E. A. Whitley, "On the interpretative flexibility of hosted erp systems," *Journal of Strategic Information Systems*, vol. 14, no. 2, pp. 167–195, 2005.

[6] S. Sahay and D. Robey, "Organizational context, social interpretation, and the implementation and consequences of geographic information systems," *Accounting, Management and Information Technologies*, vol. 6, no. 4, pp. 255–282, 1996.

[7] R. Bonney, C. B. Cooper, J. Dickinson, S. Kelling, T. Phillips, K. V. Rosenberg, and J. Shirk, "Citizen science: A developing tool for expanding science knowledge and scientific literacy," *BioScience*, vol. 59, pp. 977–984, Dec. 2009.

[8] M. Zevin, S. Coughlin, S. Bahaadini, E. Besler, N. Rohani, S. Allen, M. Cabero, K. Crowston, A. Katsaggelos, S. Larson, T. K. Lee, C. Lintott, T. Littenberg, A. Lundgren, C. Øesterlund, J. Smith, L. Trouille, and V. Kalogera, "Gravity Spy: Integrating Advanced LIGO detector characterization, machine learning, and citizen science," *Classical and Quantum Gravity*, vol. 34, no. 6, 2017.

[9] S. Bahaadini, N. Rohani, S. Coughlin, M. Zevin, V. Kalogera, and A. K. Katsaggelos, "Deep multi-view models for glitch classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2931–2935, IEEE, May 2017.

[10] M. B. Miles and A. M. Huberman, *Qualitative Data Analysis: An Expanded Sourcebook*. Sage Publications, 1994.

[11] H. K. Klein and M. D. Myers, "A set of principles for conducting and evaluating interpretive field studies in information systems," *MIS Quarterly*, pp. 67–93, 1999.

[12] N. K. Denzin and M. D. Giardina, "Introduction," in *Qualitative Inquiry—Past, Present, and Future*, pp. 9–38, Routledge, 2016.

[13] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[14] R. E. Boyatzis, *Transforming qualitative information: Thematic analysis and code development*. Sage Publications, 1998.