Chapter 3
Beyond Human-in-the-Loop: Empowering End-Users with Transparent Machine Learning

Patrick C. Shih

Abstract Advances in data analytics and human computation are transforming how researchers conduct science in domains like bioinformatics, computational social science, and digital humanities. However, data analytics requires significant programming knowledge or access to technical experts, while human computation requires in-depth knowledge of crowd management and is error-prone due to lack of scientific domain expertise. The goal of this research is to empower a broader range of scientists and end-users to conduct data analytics by adopting the End-User Development (EUD) models commonly found in today’s commercial software platforms like Microsoft Excel, Wikipedia and WordPress. These EUD platforms enable people to focus on producing content rather than struggling with a development environment and new programming syntax or relying on disciplinary non-experts for essential technical help. This research explores a similar paradigm for scientists and end-users that can be thought of as End-User Data Analytics (EUDA), or Transparent Machine Learning (TML).

3.1 Introduction

The scientific method is based on empirical measures that provide evidence for hypothesis formation and reasoning. The process typically involves “systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses” [21]. Critical thinking—“the intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, re-

P. C. Shih
Department of Informatics, School of Informatics, Computing, and Engineering
Indiana University Bloomington, Bloomington, IN, 47408, USA
e-mail: patshih@indiana.edu
flection, reasoning, or communication, as a guide to belief and action” [84]—is key to the process.

In empirical research, the scientific method typically involves a scientist collecting data based on interviews, observations, surveys, or sampling of specimens. Once the raw data sample is collected, a data cleaning and coding process identifies outliers and erroneous data resulting from sampling error, and the researcher synthesises raw data points into aggregated clusters or themes that suit the research focus. A data analytics and validation process typically follows, involving statistical or inter-coder reliability checks to ensure the quality of the findings. Finally, the results are formatted in a fashion appropriate for the intended audience, be it a research or public community. Fig. 3.1 provides a simplified view of the traditional scientific inquiry process.

![Fig. 3.1 Traditional Process of Scientific Inquiry.](image)

With the improvement of data collection instruments (e.g., space imaging for astrophysicists, environmental sampling for climate scientists, etc.) and the emergence and wide adoption of consumer Information and Communication Technologies (ICTs), researchers are turning to a broad variety of data sources to infer sample population characteristics and patterns [47]. Although improvements in data collection have enabled scientists to make more accurate generalisations and ask novel questions, the sheer amount of available data can exceed scientists’ ability to utilise or process it. Some have described this as the “Big Data” phenomena, defined by the three V’s: volume, variety, and velocity [68]. To cope, the scientific community has enlisted the help of citizen science and crowdsourcing platforms to engage the public in both data collection and data analysis [109]. However, this naturally results in a crowd management problem in which factors like task modulation, task coordination, and data verification have added to the issues that scientists must actively manage [58]. Advances in computational infrastructure and the availability of big datasets have also led to a new set of computational techniques and data analytical tools capable of processing and visualising large scale datasets [16]. This imposes a further burden on scientists, however, in the form of having to constantly learn new computational techniques and manage new visualisation tools.

Thus, crowd management and computational data analytics have become vital skillsets that the scientific workforce is starting to develop as basic building blocks of the modern day scientific method. Scientists using Big Data are increasingly dependent on knowledge of computational skillsets or on having access to technical experts in all aspects of the scientific method (e.g., data gathering, data generation, data collection, data storage, data processing, data analysis, data verification,
data representation, data sharing, data preservation, etc.). They also find themselves leveraging crowd workers who may not possess relevant scientific knowledge to provide a ground truth label of large datasets, known as “Human-in-the-Loop” (HITL) machine learning [18, 82], and scientists to correct data errors and fine-tune algorithms, known as “Interactive Machine Learning” (IML) [26, 107]. In fact, the incumbent skillsets have become so necessary and in such high demand that the White House has issued a call for a Science, Technology, Engineering, and Mathematics (STEM) initiative to make these areas of inquiry and practice more accessible to the general public [45]. Fig. 3.2 provides an overview of this newer, emerging process of scientific inquiry.

![Fig. 3.2 Emerging Process of Scientific Inquiry.](image)

Although the demand for STEM skillsets is increasing, enrolment in computer science has remained stagnant [2], which may be attributable to perceived race and gender stereotypes, or unequal access to computer science education [44, 108]. Computer education researchers have investigated how to effectively integrate computational thinking into education in order to craft, “the thought processes involved in formulating a problem and expressing its solution(s) in such a way that a computer-human or machine-can effectively carry out” [112, 111]. One approach involves motivating student interests with gamification [23, 33]. Another approach focuses on removing the technical barrier to content creation with user-friendly End-User Development (EUD) platforms [28, 63]. The latter view includes the belief that end-users with little or no technical expertise will be more willing to participate in tinkering, hacking, or other STEM activities if the barrier to entry is lowered. This research follows this second approach by proposing an end-user data analytics paradigm to broaden the population of researchers involved in this work, extending prior efforts to make computationally complex data analytics algorithms more accessible to end-users. This exploratory study focuses on examining the impact of interface design for eliciting data input from end-users as a segue into future work that will generate insights for designing end-user data analytics mechanisms.

The initial goal of this research is to create a transparent machine learning platform prototype to assist the scientists and end-users in processing and analysing real-time data streams and to understand opportunities and challenges of develop-
ing an end-user data analytics paradigm for future scientific workforces. Ultimately, the goal is to empower scientists and end-users to train supervised machine learning models to pre-process other sensor and device data streams along with those from cameras, and interactively provide feedback to improve model prediction accuracy. In this sense, the proposed end-user data analytics paradigm replaces human observers taking and coding data by hand with computational labour, where scientists or trained observers become end-users training the system by providing the system with ground truth labels for the data. In the process, the system frees the scientists from having to depend on highly technical programming expertise. In the context of a scientific workforce, this could potentially replace the onerous, labour-intensive system commonly used in observation research around the world. The same applies to domain applications with similar care and monitoring mandates, such as nursing homes, hospital intensive care units, certain security and military-related environments, and space and deep sea exploration vessels. Fig. 3.3 provides an overview of the proposed end-user data analytics paradigm.

3.2 Background

The emergence, adoption, and advances of ICTs in the past several decades have revolutionised the scientific method and the process of scientific inquiry. This section provides a general overview of the roles of ICTs in scientific inquiry along two dimensions: the scientific domain expertise of the users and the technical functions of the ICT platforms. ICT use in the scientific workforce evolved from collaboratories in the late 1980s that created communication infrastructures for scientists to share resources and early results, to citizen science platforms in the 1990s that allowed the public to contribute to scientific data collection, analysis, and interpretation. The citizen science platforms have led more recently to crowdsourcing platforms that allow online crowd workers to analyse modularised datasets (e.g., human computation and HITL machine learning). The proposed end-user data analytics platform—a transparent machine learning platform prototype that assists animal behavioural sci-
entists to analyse multi-channel high-definition video camera data—is an effort to now provide scientists with computational capabilities to process and analyse large datasets. Fig. 3.4 shows the overview of ICT use in scientific inquiry.

3.2.1 Collaboratory and Large-scale Scientific Workforce

The term “collaboratory” was coined by William Wulf while he worked for the National Science Foundation by merging the notion of traditional laboratory and collaboration that is afforded by ICT platforms that emerged in the late 1980s [60]. The shift in scientific inquiry occurred naturally out of the need to overcome physical limitations of instrument, infrastructure, and information sharing, such as results collected by scarce research instruments [1], or annotated electronic editions of 16th-century manuscripts [50]. Bos et al. (2007) describes a taxonomy of seven types of collaboratories that are differentiated by the nature of activities (loose coupling & asynchronous vs. tight coupling, synchronous) and resource needs (infrastructure and research instruments, open data, and virtual learning and knowledge communities) [9]. The early collaboratory platforms typically included functionalities such as electronic whiteboards, electronic notebooks, chatrooms, and video conferencing to facilitate effective coordination and interactions between dispersed scientists in astrophysics, physics, biology, medicine, chemistry, and the humanities [27].

3.2.2 Citizen Science

Citizen science is a two-part concept that focuses on (1) opening science and science policy processes to the public and (2) public participation in scientific projects under the direction of professional scientists [80]. Unfortunately, discussions of the public understanding of science tend to dismiss citizen expertise as uninformed or irrational, and some have advocated for involving the public in citizen projects to facilitate more sustainable development of the relationship between science, society,
and the environment [52, 53]. Although research has attempted to involve the public in citizen science projects, without proper research framing and training prior to the project, most people will not recognise scientifically relevant findings [17, 96]. Citizen science projects are also limited to those that could be broken down into modular efforts in which laypeople could reasonably participate [96]. This limits the complexity of the projects that citizens could participate in. There have been reports of mild success in terms of scientific discoveries, but the actual impact of involving citizens in scientific projects remains fairly minimal [10]. The intent in many citizen science projects is to involve volunteers in data collection or interpretation, such as the large volume of video data of animals at the zoo, that are difficult for scientists to process. These citizen science efforts are viewed as “complementary to more localized, hypothesis-driven research” [20]. Nonetheless, citizen science is generally seen as a positive factor in raising awareness of science and is frequently used as a mechanism for engaging people in civic-related projects [8, 25]. Earlier citizen science platforms typically employed traditional technologies that are commonly found in asynchronous collaboratories mentioned in the previous section [9, 27]. Modern citizen scientist platforms are starting to incorporate features found in common crowdsourcing platforms [99], and those will be described in the section below.

3.2.3 Crowdsourcing, Human Computation, and Human-in-the-Loop

Although citizen science taps into people’s intrinsic motivation to learn and contribute to science by providing labour for scientific inquiry, other crowdsourcing platforms have emerged as a way for people to outsource other kinds of labour at an affordable cost [46]. Research has linked gamification to crowdsourcing projects—if people can be incentivised to spend countless hours on playing highly interactive and engaging video games, this motivation can be harnessed as free work using progress achievement and social recognition [30, 31, 33, 85, 98]. Proponents also argue that if a task can be broken down finely enough, anyone can spend just a short moment to complete a simple task while also making a little bit of extra income. As such, crowdsourcing and human computation platforms primarily focus on task structure and worker coordination relating to workflow, task assignment, hierarchy, and quality control [58], whereas communication features between clients and workers and among workers themselves are practically nonexistent [51].

In terms of getting citizens to contribute to science projects, research has leveraged crowd workers on crowdsourcing platforms to provide ground truth label of large datasets to improve HITL prediction models [18, 26, 82, 107]. In contrast with the citizen science platforms that typically fulfil workers’ desires for educational or civic engagement activities, workers on crowdsourcing platforms are typically underpaid and have no opportunity to learn or become more engaged with the project after task completion [58]. The ethics of crowdsourcing platforms are heavily de-
bated for these reasons [42, 51, 81]. These platforms have also sparked a growth of peer-to-peer economy platforms that undercut existing worker wages [67].

### 3.2.4 From Human-in-the-Loop to Transparent Machine Learning

With the increase in available user-generated content and sensor data along with significant improvement in computing infrastructure, machine learning algorithms are being used to create prediction models that both recognise and analyse data. HITL machine learning attempts to leverage the benefits of human observation and categorisation skills as well as machine computation abilities to create better prediction models [18, 26, 82, 107]. In this approach, humans provide affordable ground truth labels while the machine creates models based on the humans’ labels that accurately categorise the observations. However, HITL machine learning suffers similar issues of crowdsourcing and citizen science platforms. For example, similar to the workers on crowdsourcing platforms, the human agents in these cases are typically used to simply complete mundane work without deriving any benefits from participation in the project. In addition, human labels suffer from errors and biases [61, 62]. Similar to the participants of the citizen science program, crowd workers are prone to making incorrect labels without domain knowledge and proper research training and framing. Accuracy in the correct identification of data and the training of the system remain two major issues in the field of HITL machine learning and machine learning as a field in general [5, 61, 62, 78]. To empower scientists with mitigating the aforementioned issues, a research agenda on an end-user data analytics paradigm is necessary for investigating issues relating to the design, implementation, and use of a transparent machine learning platform prototype to make computationally complex data analytics algorithms more accessible to end-users with little or no technical expertise.

### 3.3 Impact of Interface Design for Eliciting Data Input from End-Users

The goal of this research is to learn about the barriers that scientists and end-users face in conducting data analytics and to discover what kinds of interaction techniques and end-user technological platforms will help them overcome these barriers. As an initial step to understand current problems and practices that scientists and end-users encounter throughout the data analytics process, the following experiment was conducted to demonstrate the impact of interface design for eliciting data input from end-users.

The experiment uses NeuralTalk2 [57, 102], a deep learning image caption generator, to generate 5 most likely captions for each of 9 images. In a between-subject experiment, a total of 88 college students were randomly assigned to one of the
three interface groups—Yes/No (31 students), multiple-selection (34), and open-ended questions (23).

- In the Yes/No group, participants answered whether the generated caption accurately describes an image. This was repeated for all 5 captions for each of the 9 images, totalling 45 questions.
- In the multiple-selection group, all five captions were presented to the participants at the same time. The participants were asked to select all the captions that accurately described an image. This was repeated for all 9 images, totalling 9 questions.
- In the open-ended group, participants were asked to describe what they saw in an image. This was repeated for all 9 images, totalling 9 questions.

Participants were asked to rate their confidence level after answering each question. Participants’ feedback accuracy was assessed manually after the experiment was conducted. Selection consensus across participants and time spent were also compared and analysed. Fig. 3.5 illustrates the design of the experimental conditions. Below are results that detail how different feedback interfaces influence feedback accuracy, feedback consensus, confidence level, and time spent in how participants provide feedback to machine learning models.

Fig. 3.5 Experimental Design.

Fig. 3.6 illustrates the feedback accuracy of the captions selected by the participation. An ANOVA test followed by post-hoc comparisons revealed that the open-ended group produced higher feedback accuracy than both the Yes/No group and the multiple-selection group, and the Yes/No group outperformed the multiple-selection group (F(2,85) = 20.44, p < .0001).

Although the feedback accuracy varied significantly across groups, participants achieved similarly high within-group consensus across all 3 conditions (non-sig., see Fig. 3.7). This indicates that the differences in the feedback provided by the participants were indeed caused by the interface design conditions.

In terms of feedback confidence, although the open-ended group provided the highest level of feedback accuracy, their self-perceived confidence level (U = 372.5, p < 0.05) is as low as the multiple-selection group (U = 197.5, p < 0.01) when compared to the Yes/No group. Fig. 3.8 shows that the Yes/No group reported the highest
self-perceived confidence level. This is likely due to the fact that there leaves less room for self-doubt when the participants are presented with only Yes/No options.

Fig. 3.9 illustrates the difference in time spent for providing feedback across the 3 groups. It took the Yes/No group significantly more time to rate 45 captions (5 per 9 images) than the multiple-selection group ($F(2,85) = 6.15, p < 0.05$), whereas there is no significant difference between the open-ended and the multiple-selection groups. This is likely due to the fact that the captions in the Yes/No group were presented across a series of 45 questions instead of 9 questions presented to the multiple-selection and the open-ended groups.

Based on the results presented above, future transparent machine learning research should account for the following trade-offs when eliciting user feedback.

- The open-ended group achieved the highest level of feedback accuracy, and the participants also reported the highest level of confidence in their feedback. The fact that this can be accomplished within a similarly short time frame as the multiple-selection group points to the potential of utilising an open-ended form to elicit user feedback when the task demands such high level of accuracy. The biggest drawback is that the open-ended feedback requires active human involvement to interpret the data. A future transparent machine learning model could utilise current state-of-the-art natural language processing efforts to pre-process the open-ended responses to generate a list of possible labels before a second round of human coding. This essentially reduces the effort of analysing open-

![Fig. 3.6 Feedback Accuracy: Open-ended > Yes/No > Multiple-selection.](image)

![Fig. 3.7 Feedback Consensus.](image)
ended responses into two rounds of Yes/No or multiple-selection coding efforts for the users. The cumulative time spent in the proposed multi-round effort will not greatly exceed that of the Yes/No group based on the results demonstrated in this experiment, and the superb accuracy may justify the usage of the multi-round effort in some cases.

- While the multiple-selection group may appear to be promising due to the ease of data processing of user feedback relative to the open-ended group, the results show that it produced the lowest feedback accuracy and the participants are less confident of their feedback. One advantage of this user feedback elicitation method is that it gives the users the ability to view and provide feedback on multiple machine-generated labels at the same time, which results in the lowest cumulative time spent for the participants in our experiment. This method may be desirable in situations where feedback accuracy is less critical and the goal is to process through a large amount of data in a short period of time.

- The Yes/No group produced the medium level of feedback accuracy. Although the participants in the Yes/No group spent the highest cumulative time to provide feedback, it took the participants much less time to rate the individual options with the highest self-reported confidence level compared to the multiple-selection and the open-ended groups. The flexibility of adjusting the number of options that the users rate at any given time (e.g., users can stop after rating through 2 options instead of having to view all of the options at once in the
multiple-selection group) can be especially desirable when user commitment is
unknown and the intention is to minimise user burden to provide feedback. The
human-labelled results are also easy for machine learning models to process,
making the Yes/No group the most flexible and adaptable method.

These experimental findings show that interface design significantly affects how
end-users transform information from raw data into codified data that can be pro-
cessed using data analytics tools, and the insights can inform the design, imple-
mentation, and evaluation of a usable transparent machine learning platform in the
future. Future transparent machine learning could expand the study to different user
feedback scenarios and contexts that require human feedback.

3.4 Design for End-User Data Analytics

Currently, there are many popular, general-purpose open-source scientific numerical
computation software libraries such as NumPy [103], Matplotlib [48], and Pandas
[70] that users can import into their software development environment to conduct
numerical analysis programmatically. However, the use of these software libraries
requires significant programming knowledge. To make data analytics more user-
friendly, popular machine learning and data mining software suites such as Weka
[32, 113], Orange [19], KNIME [7], and Caffe [56] provide users with command-
line and/or graphical user interfaces to access a collection of visualisation tools and
algorithms for data analysis and predictive modelling. Yet these software suites do
not provide label suggestions based on the currently trained model, typically oper-
ating under the assumption that ground truth labels are error-free. Functionalities
of these software suites are typically limited to training static rather than real-time
live-stream datasets and lack the ability to allow users to interactively train ma-
chine learning models in order to more effectively explore data trends and correct
label errors. In other words, these platforms neglect data collection and data (pre-
)processing phases, both of which are essential steps throughout data analytics. A
new paradigm is needed to more effectively disseminate the data science mindset
more holistically and make data analytics more accessible to learners and end-users.

To realise intuitive, easy-to-learn, and user-friendly interfaces for data collection,
processing, and analytics, it is necessary to create a series of software front-end
prototypes, increasing in complexity but all sharing the same basic framework for
interaction. The goal of the prototypes will be to learn about how different interac-
tion techniques can replace or enhance the current paradigm of data processing by
scientists and end-users. In the spirit of end-user development paradigms such as
Scratch [79], combining interaction techniques used in interactive machine learning
[26, 107] and direct manipulation interfaces [49] to create a novel interface to ease
the training process of supervised learning models could potentially yield a more
usable transparent machine learning platform. The goal is to create a system that
allows the user to smoothly move between data and a list of inferred behaviours,
allowing scientists and end-users to visually preview and make corrections to the
prediction model. Although the prototypes will vary, the interactions will share the same basic features. Users will use the platform to select the input data streams to be worked and then overlay these with behavioural data previously coded by trained scientists and end-users.

### 3.5 Conclusion

Successful investigations of transparent machine learning require multidisciplinary expertise in (1) human-computer interaction and end-user oriented design processes such as participatory design, interaction design, and scenario-based design [12, 14, 59, 64, 83, 92, 95, 97, 104, 3, 4, 66, 65, 24, 29, 74], (2) human computation and crowdsourcing [6, 37, 86, 100, 105, 106, 114, 38, 22, 75, 11, 36, 40, 41, 13], (3) end-user visualisation interfaces and computational data analytics [35, 39, 54, 55, 73, 88, 89, 91, 93, 110, 116, 101, 72, 71, 43, 34, 87, 94], and (4) computer science education [15, 76, 77, 115, 117, 118, 69]. This research reveals the initial insights on how to make data analytics more accessible to end-users, to empower researchers in scientific inquiry, and to involve the public in citizen science. This research also will provide trained end-users opportunities to participate in citizen science efforts, allowing them to contribute directly to citizen science as well as become more familiar with the scientific method and data literacy, heightening awareness of how STEM impacts the world.

There are numerous potential applications of this work. Sensor and surveillance technologies have made great strides in behaviour profiling and behavioural anomaly detection. Such technologies may allow scientists and end-users to closely observe real-time data streams around the clock. Although the proposed end-user data analytic and transparent machine learning platform is currently targeted toward scientists and end-users, the platform and the resulting knowledge could be used most immediately to make data analytics more accessible for other domain applications with similar care and monitoring mandates, such as nursing homes, hospital intensive care units, certain security and military-related environments, and space and deep sea exploration vessels.

### References

45. House, W.: President obama to announce major expansion of educate to innovate campaign to improve science, technology, engineering and math (stem) education. Office of the Press Secretary (2010)