Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes

Tom Hitron
Media Innovation Lab (miLAB)
Interdisciplinary Center
Herzliya, Israel
tom.hitron@milab.idc.ac.il

Yoav Orlev
Media Innovation Lab (miLAB)
Interdisciplinary Center
Herzliya, Israel
yoav.orlev@milab.idc.ac.il

Iddo Wald
Media Innovation Lab (miLAB)
Interdisciplinary Center
Herzliya, Israel
iddo.wald@milab.idc.ac.il

Ariel Shamir
Interdisciplinary Center
School of Computer Science
Herzliya, Israel
arik@idc.ac.il

Hadas Erel
Media Innovation Lab (miLAB)
Interdisciplinary Center
Herzliya, Israel
hadas.erel@milab.idc.ac.il

Oren Zuckerman
Media Innovation Lab (miLAB)
Interdisciplinary Center
Herzliya, Israel
orenz@idc.ac.il

ABSTRACT

Machine Learning services are integrated into various aspects of everyday life. Their underlying processes are typically black-boxed to increase ease-of-use. Consequently, children lack the opportunity to explore such processes and develop essential mental models. We present a gesture recognition research platform, designed to support learning from experience by uncovering Machine Learning building blocks: Data Labeling and Evaluation. Children used the platform to perform physical gestures, iterating between sampling and evaluation. Their understanding was tested in a pre/post experimental design, in three conditions: learning activity uncovering Data Labeling only, Evaluation only, or both. Our findings show that both building blocks are imperative to enhance children's understanding of basic Machine Learning concepts. Children were able to apply their new knowledge to everyday life context, including personally meaningful applications. We conclude that children's interaction with uncovered black boxes of Machine Learning contributes to a better understanding of the world around them.

CHI 2019, May 4–9, 2019, Glasgow, Scotland Uk

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300645

CCS CONCEPTS

• **Human-centered computing → User interface management systems**; • **Applied computing → Interactive learning environments**;

KEYWORDS

Machine Learning, Children, Learning System, Construction kits, Design principles

ACM Reference Format:
Tom Hitron, Yoav Orlev, Iddo Wald, Ariel Shamir, Hadas Erel, and Oren Zuckerman. 2019. Can Children Understand Machine Learning Concepts? The Effect of Uncovering Black Boxes. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland Uk.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3290605.3300645

1 INTRODUCTION

Machine Learning (ML) processes are integrated into products and services that influence our everyday lives, changing the way people interact with technology. ML allows computing systems to learn directly from examples, data, and experiences, and has the potential to become a transformative technology [33]. However, the underlying processes of ML are rarely exposed to users and are not intuitively understood. The vast development of ML in industry and academia is expected to further expand ML products integration, while novices lack opportunities to acquire accurate ML mental models. ML learning activities are still scarce, unlike other computational concepts that are introduced to novices through coding classes and making activities. ML processes are not similar to the standard set of computational concepts novices are exposed to when learning coding, and require a dedicated learning activity. Hence, understanding basic ML concepts is becoming important for people of all ages,

Figure 1: A girl training the ML system using the input device (child photographed with permission).

including children, who are growing up in an environment that integrates ML products more than ever before.

Children constantly learn from experience by interacting with the physical world around them [7, 28, 39]. This direct exploration contributes to the construction of mental models, which are conceptual and operational representations of phenomena and processes in the world [18]. Direct exploration is limited when processes are "black-boxed" (i.e. hidden from the user), making it harder to construct accurate mental models [3]. Commercial digital services are often designed to black-box complex processes, in an attempt to increase the product's ease of use and consumers' adoption [15]. Children's interaction with black-boxed processes may lead to the development of inaccurate or oversimplified mental models [16]. Once formed, these inaccurate models become difficult to overcome. Therefore, exposing children to non-black-boxed processes is imperative to the formation of accurate mental models [15]. The effect of uncovering black boxes in the context of ML was demonstrated with adults useres, showing better understanding of ML when black boxed processes were uncovered [23]. This has yet to be tested with children. It is important to note, however, that uncovering too many processes can interfere with the learning process as a novice learner may be overwhelmed [31, 32]. Therefore, striking the balance between black-boxed processes and uncovered processes is of great importance. Today many children interact with ML products and services such as natural language processing (Amazon Echo, Google Home) [24] or face recognition (Snapchat filters) but are not exposed to their underlying processes. Such black-boxed experiences may lead to inaccurate or oversimplified mental models of ML.

In this work we focused on supervised ML, one of the three key branches of machine learning, where a system is trained with labelled data [33]. There are several approaches to define the underlying processes of supervised ML [22]. As this paper is aimed at introducing ML processes to children, we focus on classification problems, that are relatively less complex and are common in real world applications. In these problems, examples are labeled into classes, and are then used to train a model that is able to classify new examples [1]. This type of supervised ML can be defined as a pipeline consisting of four building blocks [26]: Data Labeling, Feature Extraction, Model Selection and Validation, and Evaluation. (1) Data Labeling (or Gathering): collecting data points and the classification of each one to a dedicated category. The training process requires sufficient amount of data for a ML model to recognize patterns, a clear classification of what is considered a positive example for the class and negative examples for the class (examples not included in the class), and a sufficient variety of data that represents the different examples within the class, including instances close to the boundary between positive and negative examples [1]. For the purpose of this paper, we term these requirements as "Data Labeling Aspects" and simplify their definition to: Sample Size (sufficient amount of data); Sample Versatility (sufficient variety of positive examples within a class); and Negative Examples (inclusion of examples that do not belong to a class). (2) Feature Extraction: Preprocessing the data in order to simplify classification and speed up computation. This includes finding the useful features that are fast to compute rather than feeding the raw data to the algorithm [5]. (3) Model Selection and Validation: The consideration and testing of different model types (and their parameters) in order to find the best one for a particular application [5]. (4) Evaluation: testing the trained ML model with new data to evaluate the quality of the chosen model [1].

Some of these ML building blocks are more accessible than others. Feature Extraction and Model Selection are more complex and harder to understand for novices, while Data Labeling and Evaluation have been suggested to be more accessible [38]. Therefore, in this work we uncover Data Labeling and Evaluation building blocks and black-box the other two. It is, however, not clear whether children are able to comprehend even the more accessible concepts of ML. It was previously believed that children need to reach certain maturity in order to comprehend complex concepts [6]. However, there are consistent indications for children's ability to understand complex concepts through iterative experimentation of trial and error [27] in several domains, including probability [42], systems thinking [43], kinematics [34], and AI [8, 9].

To test if direct experience with accessible ML building blocks contribute to the understanding of core ML concepts we designed an interactive learning system and evaluated it with 30 children. The system provides an opportunity for learning through direct experience and iterative exploration

of the Data Labeling and Evaluation ML building blocks (see Figure 1). We assessed children's ability to understand ML concepts, apply their understanding to a new context, and generate accurate examples for new ML applications relevant to their daily lives.

Previous work includes systems designed to promote children's hands-on experience with ML processes, and systems designed to promote children's learning of other complex concepts.

Several projects have introduced systems for hands-on experience with ML concepts without evaluating their effect on ML understanding. Snap! are AI visual programming blocks for children [19], enabling creation of ML applications using pre-existing models. In industry, Google introduced two projects, the AIY kit encourage children to build a homemade smart speaker, however the underlying ML processes are black-boxed; the Teachable Machine is an image recognition system for a computer vision algorithm that evaluates the system's ability to identify new examples.

Systems designed for children's learning of complex concepts, but not ML processes, involve hands-on experiences and promote iterations. Within the tangible interface research community, Flow Blocks are an example for a system designed to promote the understanding of complex systems [42]. Through direct experience children gain an understanding of probability and dynamic behavior. Within the construction kit research community, kits were designed to develop children's understanding of abstract concepts in mathematics, science, and engineering, including roBlocks for kinematics and distributed control [34]; WayMaker for map topology; and DemBones for balance in motion [36].

The Learning By Design (LBD) approach [21, 32] argue that being engaged in design and modeling enhances the learning of complex systems through systematic exploration. Hmelo-Silver (2000) showed that when children design artificial lungs and build partial working models of the system, they develop an understanding of the human's respiratory system [16].

Only a few studies addressed learning of complex processes in the ML and AI domain. Druga et al. (2018) evaluated children's perception of a robot's capabilities [9], suggesting that hands-on experience of navigating a robot may refine children's understanding of the the robot's AI processes that control navigation. Woodward et al. (2018) conducted a co-design study with children, showing that participation in the co-design process of a new intelligent user interfaces enabled children to conceptualize and propose ideas for complex technical systems that require artificial intelligence processes [41].

Our recent Work-in-progress looked more specifically into children's learning of ML concepts [14]. In that preliminary work, we conducted a Wizard-of-Oz experience designed to give children feedback when "training" a device. Our initial findings showed that a direct experience with accessible ML building blocks have the potential to enhance children's understanding of basic ML concepts.

We extend prior work by implementing a ML gesture recognition system that uncovers the more accessible ML building blocks of Data Labeling and Evaluation, using a physical input device with an embedded acceleration sensor. The ML system was specifically structured to promote children's learning through design, based on principles previously indicated as an effective learning method for complex concepts [21]. Our learning system enabled children to collect data, design a ML model, and revise the model based on feedback.

2 SYSTEM DESIGN AND IMPLEMENTATION

We implemented 'Gest', a ML gesture recognition system, to be used as a research platform for studying children's learning of ML concepts. The choice of hand gesture recognition was grounded in the "noisy" nature of physical hand gestures, that are rarely similar to one another. The "noisy data" sample requires a thorough training process involving frequent iterations, making the sampling processes similar to real-world "noisy" data collection processes. In addition, physical hand gestures are common in children's physical play activities and involve a hands-on experience, previously shown to increase motivation, interest, and engagement [17, 29].

The Gest system consists of three components (see Figure 3): our previously published hardware device with an embedded acceleration sensor for data collection [13]; a software module for data analysis that uses GRT (an open source gesture recognition toolkit); and a simple interface for control and feedback. The input device is used for Data Labeling by collecting hand gesture samples. Data Labeling was done using the interface, that was also used to transfer the labeled samples to the GRT module for algorithm training (see Figure 2). The Evaluation building block was applied by a recognition evaluation process: new gesture samples were classified by the model and real-time feedback was presented using the interface, allowing users to quickly assess recognition accuracy. In addition, the system was designed to facilitate easy transition between the Data Labeling and Evaluation phases.

Design principles for Gest: a ML learning environment for children

Based on prior work in the constructivism school of thought and in cognitive psychology, we identified a set of principles for guiding children's learning.

Design Principle 1: Low Floor. Construction kit literature emphasizes the Low Floor principle to encourage learning by

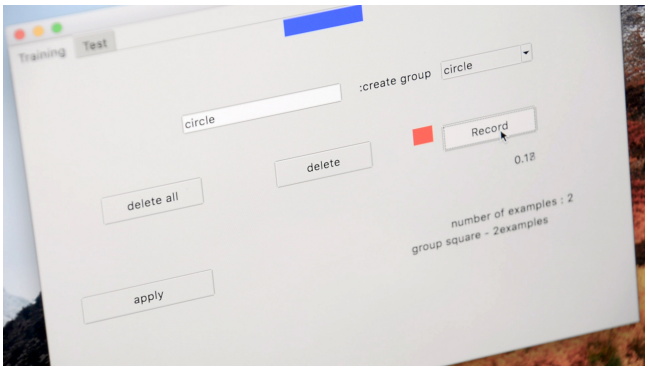

Figure 2: The system's interface Data Labeling screen, enabling children to record new data samples, delete them, label the samples, create new class, and apply the labeled sam-

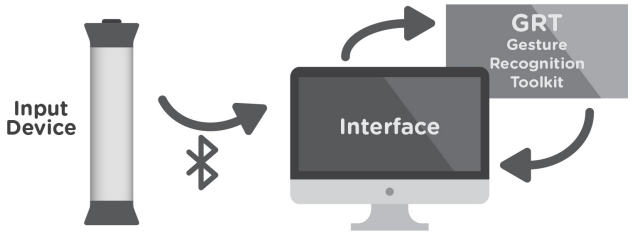

Figure 3: The system components include an Input Device with embedded accelerometer, a GRT software module for data training and analysis, and an interface to provide control and feedback to the user.

doing. Low Floor is defined as a learning experience that does not require any prior formal knowledge and allows immediate exploration. This can be achieved by implementing a small number of features, that are simple and specific, promoting quick understanding that empower children to explore the system without barriers [32].

Design Principle 2: Uncovering black-boxes. Uncovering black boxes can promote direct experience with underlying processes, but may also introduce complexity that will limit learning. Prior work suggested to carefully choose which black-boxes to uncover [32] in order to strike the right balance between promoting learning of selected processes (by uncovering the black boxes of these specific processes) and maintaining an accessible and understandable experience (by keeping other processes black-boxed).

Design Principle 3: Promote Iterations. Researchers from the constructivist school of thought suggest that iterations, debugging, real-time feedback, and reflection promote learning [11, 37, 40]. They emphasize that learning is commonly formed when the course of action is modified or changed based on continuous cycles of trial and error that involves reflection [40]. Reflection is thought to initiate a debugging cycle that involves thinking on the problem and generating possible solutions [37].

Design Principle 4: Promote Self-generated Knowledge. This principle is based on the Generation Effect from cognitive psychology literature, indicating that learning processes that provide learners with the opportunity to generate knowledge based on their own experience, strengthen memory traces of the gained knowledge and facilitate information recall [12]. This type of learning, defined as the Generation Effect, was shown to create stronger, long lasting knowledge in comparison to learning based on information that is presented to the learner in a passive manner, including reading, listening to a source of information, or even memorizing by mere repetition [4, 35]. This effect is explained by the deeper cognitive processing required for information generation [25]. In addition, constructivist research suggests that when there is a conflict between new self-generated knowledge and previous knowledge, the resolution of such cognitive conflict leads to increased understanding [20]. Learning systems can leverage this robust phenomenon through design, and increase the opportunities for self-generation of information.

Building on these four principles, the Gest system was designed as follows:

- To create a "Low Floor" experience we developed a system with a limited set of features, that are simple and specific, enabling users with no prior knowledge to immediately explore it.
- For Uncovering black-boxes, we focused on two ML building blocks that are considered more accessible than others, keeping Feature Extraction and Model Selection building blocks as black-boxes, and uncovering Data Labeling and Evaluation building blocks.
- To promote iterations we implemented a simple one-button interface for children to move back-and-forth between the Evaluation and Data Labeling phases, allowing children to effortlessly revise their data based on evaluation feedback.
- To promote self-generation of information we refrained from an explicit explanation of the ML processes. Instead we designed a system for direct experience with the uncovered ML building blocks. The children, we provided with a system allowing to perform Data Labeling by themselves. The system was accompanied by three structured learning tasks that promoted cycles of trial-and-error related to different Data Labeling Aspects (see Figure 4).

Technical Implementation

The Gest system consists of three components: an input device, a software module, and an interface (see Figure 3).

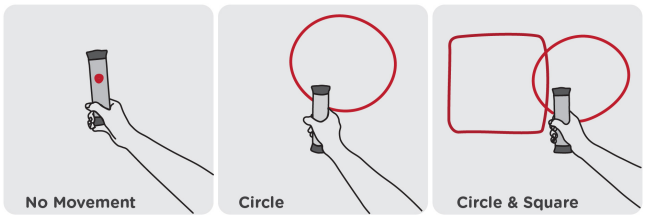

Figure 4: Illustration of the three learning tasks. Train the device to recognize: 'no movement'; 'a circle'; 'a circle and a square' (from left to right).

The hardware device we used is a previously-published stick-like digital device, designed specifically for children, with a plastic case that affords holding [13]. We extended the existing input device with Arduino code that transmitted the desired sensor data via Bluetooth to the data analysis software on a nearby laptop. Python code received the data, processed it and transferred it for gesture recognition using ML analysis (training or classification). For gesture recognition, we used the open-source Gesture Recognition Toolkit (GRT), developed by Gillian and Paradiso [10].

ML Process as Black Box 1: Feature Extraction. The input device sensors included three types of movement sensing with 3 DoF each: accelerometer, gyroscope and magnetometer. Therefore, the sensor data stream had 9 features. We tested and analyzed many subsets of these 9 features in order to find the most appropriate set for the hand gesture recognition tasks in our study. We empirically evaluated many different combinations of sensing features, and found that the gyroscope's Angular Velocity had the best recognition accuracy.

ML Process as Black Box 2: Model Selection and Validation. For the classification task, we used the Dynamic Time Warping (DTW) algorithm. The DTW is a time series analysis algorithm that compares two sequences. The DTW identifies the class with the highest probability, meaning it will always classify a series in one of the possible classes. While the GRT enabled some parameter tuning such as Warping Radius, once we isolated the Angular Velocity features we learned that the default setting of 0.2 Warping Radius provided the best results.

3 USER STUDY

We tested whether a learning experience with the Gest system promotes children's understanding of ML concepts. Specifically, we evaluated whether children who trained the input device to identify gestures, understood the Data Labeling Aspects: Sample Size, Sample Versatility, and Negative Examples. We compared children's learning in three conditions: a Full System, uncovering both Data Labeling and Evaluation building blocks, a Partial System uncovering only Data Labeling, and Partial System uncovering only Evaluation. This comparison allowed to assess if uncovering both building blocks is essential for learning ML concepts, or if one of them is sufficient. While the baseline conditions do not simulate everyday interaction with ML, they allow to evaluate which building blocks are essential for learning.

To provide further insight towards the extent of the learning effect with the full system, we conducted an additional evaluation. At the end of the learning experience children in the Full System condition were asked to apply the ML concepts they learned to everyday situations, and to generate new ideas involving ML processes in the context of their daily lives.

Method

Participants. 30 children participated in the study (20 boys and 10 girls, age range 10-13, M = 11.59 SD = 0.97), recruited from a local children coding event (Scratch event) and through personal acquaintances with the researchers. Participant's experience in technology varied from basic (smartphone apps and gaming) to advanced (experienced with coding). We followed ethics guidelines including IRB, parental consent, children consent, and parental approval for pictures and videos. In addition, we followed Read's (2015) guidelines for research with children [30]. All children that participated in the research were invited to a guided tour in the author's research lab, followed by a 3D printing activity.

Experimental design. We applied a mixed experimental design that included a within-participant pretest vs. posttest evaluation of ML concepts understanding, and a between-participant comparison of three conditions, each with different ML building blocks: both Data Labeling and Evaluation; only Data Labeling; and only Evaluation (see Figure 5).

Dependent measures. To evaluate understanding of ML concepts, children were given examples for ML applications and were asked to explain the underlying processes. We assessed

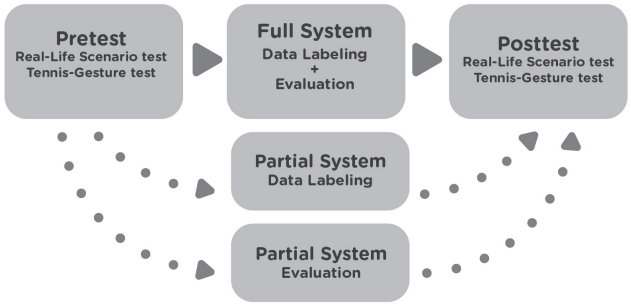

Figure 5: The mixed experimental design included a within-participant pretest vs. posttest of ML concepts understanding, and a between participant comparison of three conditions.

two types of ML applications: (1) Tennis gestures recognition: new ML examples in a context similar to the learning experience (hand gestures context), one example in the pretest and one in the posttest; and (2) Real-life ML application: ML examples in a context different than the learning experience, one in the pretest and one in the posttest. The examples were counterbalanced (pretest/posttest) between participants.

Same context examples (Tennis Gestures test): The system was trained by the researcher to identify a Forehand and Backhand tennis gestures. In the test, children were asked to perform the gestures by themselves with the input device, observe the system's feedback and explain the gesture recognition process (i.e "How does it work?"). Children's explanations allowed us to assess their understanding. Forehand and Backhand gestures were counterbalanced between participants (pretest/posttest).

Different context examples (Real-life Application Scenarios test): Two scenarios of Real-Life ML applications were presented. A "smart speaker with speech recognition" and, an "autonomous car image recognition". In the test, children were asked to explain the recognition process (i.e "How does it work?"). The scenarios were counterbalanced between participants.

- Smart speaker with speech recognition: "A family bought a new internet-enabled system and placed it in the living room. When one of the family members wanted to hear a song, they would say the word 'Song' and the system would recognize the person's voice and play that family member's favorite song."
- Autonomous car image recognition: "Dan bought a new autonomous car. The car can drive and navigate independently. The car can recognize road signs and obstacles on the road and respond accordingly. For instance, in the event that a child suddenly crosses the road, the car can detect it and stop immediately, however if the car detects a traffic light with a green light, it will continue driving."

Participants in all groups performed three phases: (1) a pretest to assess their understanding prior to the learning experience; (2) the learning experience with the system according to their condition; and (3) a posttest to assess their understanding after the experience. Participants in the Full System condition also participated in a semi-structured interview to assess general understanding of ML concepts (see Figure 5).

Participants were told they are helping the research team test a new product. The study started with a general question about the children's prior knowledge of ML and Artificial Intelligence (AI) - "Have you ever heard the term ML or AI"?, "Can you share what you already know about ML and AI"? All children stated they are not familiar with ML or AI. After this initial assessment, the pretest began.

The pretest was followed by the learning experience with the system in one of the three conditions. Children were presented with the input devices and informed that it includes a movement sensor. In the Full System condition, children were told that the system can detect their movement and that after performing each gesture they should classify it into a category they define using the interface (see Figure 2). The researcher did not provide any verbal explanation regarding Data Labeling Aspects. Children were also informed that when they believe the sample is sufficient they should progress to the Evaluation phase by pressing the "apply" button. In the Evaluation phase, children were able to test the system's recognition accuracy with new gestures, and if they wanted to they could click a button and return to the Data Labeling phase to improve the training.

Learning Tasks. The learning experience was guided by three learning tasks designed to provide children with the opportunity to generate their own understanding of the three Data Labeling Aspects (Sample Size, Negative Examples, Examples Versatility). Each learning task involved two phases: first, children were asked to sample and label gestures until they believe the sample is sufficient. Children were then invited to observe the system's real-time feedback when performing new gestures. Children were notified that an accurate recognition means the feedback presented corresponds with the gesture performed by the child. The three learning tasks instructions were: "Train the device to recognize a 'no movement' gesture"; Train the device to recognize a 'Circle gesture'; Train the device to recognize a 'Circle and Square' gesture (see Figure 4). In the *No movement* learning task children were requested to train the system to recognize states in which the input device is not moving. To accomplish this task, children had to understand the need to create a group consisting of examples where the device does not move. During this process, children learned that they also need to create a group consisting of examples where the device is moving. In the *Circle* task children were asked to train the system to recognize a circle movement. To accomplish this task, children had to figure out that they need to sample several examples of circle gestures and label them as belonging to the circle class, and to create a negative class for no-gesture or gestures that are not a circle. In the *Circle and Square* task children were asked to recognize both square and circle. To accomplish this task, children had to figure out that they need to give examples for different types of circles and different types of squares, as well as negative examples.

At the end of the learning experience children were asked to explain the process they went through. They were asked: "How would you explain the activity to a friend ?". Following

that question, children performed the posttest that included the "same context" and "different context" tests. Following the posttest, children in the Full System condition were asked two further questions about their perception of ML in their daily lives. The first question was general: "How would you use ML technology in your daily life?"; and the second was an ethical question, regarding possible risks of ML: "Are there cases where we should not use ML?". All sessions were documented using a video camera for further analysis.

The two Partial System conditions were identical to the Full System condition, apart from the exclusion of either Data Labeling or Evaluation phases. In the Data Labeling only condition children were not asked to evaluate the accuracy of their training and did not receive feedback. In the Evaluation only condition children were notified that we trained the device to recognize certain movements: a circle gesture, a square gesture, and "no movement". They were asked to evaluate the system's recognition accuracy but were not asked to sample and label gestures. In line with the ethics guidelines for research with children, after completing the experiment, children in both Partial System conditions performed the Full System activity [2].

Researcher's structured support. To support children that had significant challenges progressing in the learning experience, the researcher provided pre-defined structured support that was applied in two specific cases: if the child did not understand the system's functions, or if the child misinterpreted the system's feedback. Specific sentences were selected to support each case. For the first case: "The system works exactly according to how you taught it in the training phase."; "The device knows only what you trained it to know." For the second case: "Are you sure that the system has recognized the movement correctly?"; "What would you do to check if the system isn't working correctly?" Beyond these predefined sentences, the researcher was not involved in the children's self-learning process.

4 DATA ANALYSIS

Data analysis included video coding, interview transcriptions, and a two-way ANOVA analysis. A primary coder coded all videos and interviews per participant and a second coder coded 50% of the videos and interviews independently. Interrater reliability was found to be high (Kappa = 92%). Children's explanations of the ML applications in the pretest and posttest were analyzed to evaluate understanding of ML concepts. We assessed whether children did or did not use the different Data Labeling Aspects in their explanations. For example, a child mentioning that the device should be trained with multiple examples, was considered as understanding the need for a large Sample Size. A child mentioning that the device needs to be trained with various examples, was an indication for understanding the need for Sample Versatility. A child mentioning that the device needs to be trained with different examples than the target examples, was considered as understanding the need for Negative Examples. After this analysis process, we calculated the differences between the number of Data Labeling Aspects children understood in the posttest (range of 0-3) and the number of Data Labeling Aspects they understood in the pretest (range of 0-3). The difference was used to evaluate children's general learning effect. This analysis was performed twice for the two dependent measures: (1) tennis gestures - explanations of ML applications in a similar context to the learning context, and (2) real-life scenarios - explanations of ML applications in a different context. We performed the two-way ANOVA to evaluate the influence of the type of interaction with the system (Full System and the two Partial System conditions) on the number of Data Labeling Aspects understood in the learning experience (Number of Data Labeling Aspects understood in the pretest vs. the Number of Data Labeling Aspects understood in the posttest). During the final interview, children gave examples for applying ML in their daily lives. We analyzed their answers and categorized them into three predefined themes: accurate ML examples (examples that require ML process for a successful operation, e.g. data collection is required to develop a model to identify a situation); non-ML examples (Examples that do not require ML process for a successful operation, but require other (possibly simpler) computational processes, for example a sensor controlled with a simple if/else condition); and fictional examples (Examples that require highly complex or non-existing technologies for successful operation, for example remotely detecting electricity patterns in the human body).

5 FINDINGS

Findings include children's explanations for the Data Labeling Aspects in the learning experience, two-way ANOVA comparisons, and a qualitative analysis of children's responses in the pretest and posttest (see Table 1 for the number of children who understood each principle during the learning experience). In addition we present a qualitative analysis of the interviews, which includes examples of children's ideas for new ML applications. All children's quotes were translated to English from their original language.

Data Labeling Aspects: in the learning experience

The analysis of children's answers to the question at the end of the learning experience ("How would you explain to a friend what you did?") revealed that all the children in the Full System condition understood the Data Labeling Aspects in the context of the learning experience. Children explained that they had to provide the system with a large sample (verifying Sample Size): "I showed many many examples of

circles, and then it learned, learned from me" (p.16). They stated that the examples had to be versatile (verifying Sample Versatility): "I've shown it examples which are different than what I would typically do, like larger ones, smaller ones (circles)... to give it more opportunities to learn" (p.6). Children also explained the need for Negative Examples: "I wanted to show the device what is moving as opposed to not moving" (p.25).

Data Labeling Aspects: in the Same Context

A two-way ANOVA revealed that the type of interaction with the system (Full System and the two Partial System conditions) significantly influenced the number of Data Labeling Aspects children were able to understand during the learning experience, as indicated by the significant interaction between the system conditions and pretest vs. posttest principles' understanding [$F(2,27)=16.19$, $p<0.01$]. Post-hoc multiple comparisons using Scheffe's method on the system conditions revealed that the Full System condition was the only condition that resulted in improved understanding [i.e. an increase in the number of Data Labeling Aspects used by the children in the posttest compared to the pretest; $t(1,27)=5.07$, $p=0.008$; $t(1,27)=4.76$, $p<0.001$]. No difference was found between the two Partial System conditions (see Figure 6).

In the pretest, most of the explanations children gave for the Same Context (tennis-gesture) application were inaccurate in all conditions, and rarely included any relation to the Data Labeling Aspects. Some of the answers included non ML-related technical explanations for gesture recognition: "It detects it according to the strength of the swing" (p.24); "It can recognize when it moves up or down" (p.8). Other answers were inaccurate or fictional: "I think it can recognize the air that goes inside and can calculate the speed" (p.18). Some children stated they cannot explain the processes as they do not have enough knowledge. Only three children gave valid explanations in the Same Context pretest: "You have a movement sensor inside, and they showed the device many times how a forehand looks like and the system just learned" (p.16). In the posttest, more children were able to provide accurate explanations. This change was significant only in the Full System condition: "They showed the device multiple examples of backhand, and multiple examples of 'not moving', or of doing things which are not backhand, but, if you do a forehand it might detect that you are not moving or doing backhand, this is why you have to record many different examples" (p.16); "The device was given many examples of backhand, but also examples of not moving, and other gestures too. You have to show a variety of backhands so it could get it right" (p.6); "I suppose it's the same as I did in the activity, you recorded many backhands" (p.8). In the two Partial System conditions, most children struggled with providing an accurate explanation, in a similar way to the pretest: "[The device] works with electricity and sensors"(p.4); "[The device] has a GPS inside to detect its location" (p.29).

Data Labeling Aspects: in a Different Context

A two-way ANOVA revealed that the type of interaction with the system (Full System and the two Partial System conditions) significantly influenced the number of Data Labeling Aspects children were able to understand during the learning experience, as indicated by the significant interaction between the system conditions and pretest vs. posttest principles' understanding [$F(2,27)=8.46$, $p<0.001$]. Post-hoc multiple comparisons using Scheffe's method on the system conditions revealed that only the Full System condition resulted in improved understanding [i.e. an increase in the number of Data Labeling Aspects used by the children in the posttest compared to the pretest; $t(1,27)=3.38$, $p=0.008$;

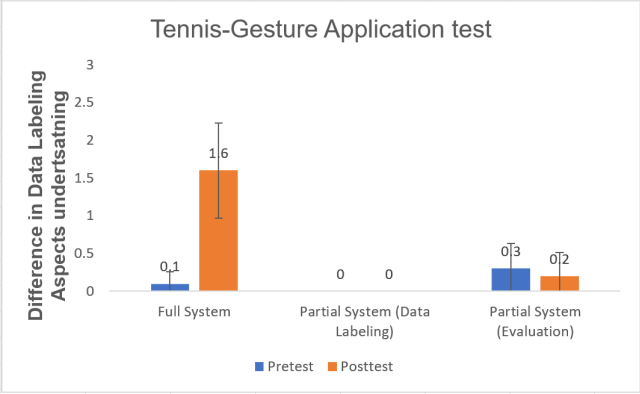

Figure 6: Difference between children's understanding of Data Labeling Aspects before and after the learning experience (Pretest subtracted from Posttest) in the same context test.

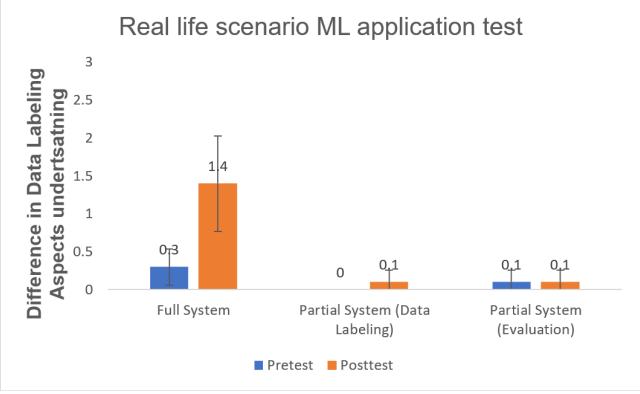

Figure 7: Difference between children's understanding of Data Labeling Aspects before and after the learning experience (Pretest subtracted from Posttest) in the different context test.

Table 1: Number of children who stated each data labeling aspect in their explanation

		Real life scenario ML application test		Tennis-gesture application test	
		pretest	posttest	pretest	posttest
Sample Size	Full System	3	6	1	7
	Partial System: Data Labeling	0	1	0	0
	Partial System: Evaluation	1	1	2	1
Sample Versatility	Full system	0	5	0	6
	Partial System: Data Labeling	0	0	0	0
	Partial System: Evaluation	0	0	1	1
Negative Examples	Full system	0	3	0	3
	Partial System: Data Labeling	0	0	0	0
	Partial System: Evaluation	0	0	0	0

t(1,27)=3.72, p=0.005]. No difference was found between the two Partial System conditions (see Figure 7).

In the pretest, most of the explanations to the Different Context application (real-life scenarios: Autonomous car or Smart speaker) were inaccurate in all conditions, and rarely included any relation to the Data Labeling Aspects. Some children suggested non-ML technical explanations for the scenarios: "The car has lasers that recognize if there is something in front of the car, if the laser hits something, the car stops" (p.18). Other children provided explanations that were inaccurate or fictional: "The car can detect the electricity in the human body and respond accordingly" (p.21). Some children declared that they do not know how to explain the processes. Only two children provided an accurate explanation during the pretest: "The car has 360 degrees cameras, and you give it many examples of children and than you kind of keep doing the same thing" (p.16); "The car has huge databases of photos that somebody inserted, then the car checks what it sees compared to the database" (p.19).

In the posttest, children that participated in the Full System condition gave more accurate explanations: "People from the family had to speak to the device, saying the same word many times, in different tones" (p.14); "When creating the car you have to define for it multiple and different examples of children" (p.7). The learning experience with the Partial System did not lead to improved understanding and most children provided inaccurate explanation: "It probably has lasers that can detect the height of the person standing in front of the car" (p.27); "The car has a sensor and it was programmed to detect things like DNA and photosynthesis" (p.30).

Interview analysis: daily life ML applications

80% of children that participated in the Full System condition generated at least one accurate example of ML application relevant to their daily lives: "I take swimming classes. I could use it for counting, I will teach it to recognize only specific movements so it could follow the different swimming styles I do in the pool. I would also want something that could monitor my concentration level in class, according to my facial expressions" (p.7); "It could be used in supermarkets, so the cashier could recognize Broccoli without scanning a barcode" (p.1); "I am watching a dog for a while now, I would really like something that could tell me about her needs, maybe according to her movements, like when she wants to go for a walk or when she is hungry" (p.6). The two children that did not generate an accurate example, gave sensor-related examples that do not leverage ML processes "I would like something that could help me feed my bird, maybe when it feels that the bowl is empty, something that can recognize the weight of the bowl" (p.14); one example was fictional: "a robot that will come every morning to fix my bag for school and make me breakfast" (p.8). When asked if there are areas where ML should not be used, 50% of the children described relevant scenarios. They mentioned safety issues: "Computers will always make mistakes, you can't trust them entirely, it's like when I trained the device and it did not recognize square or circle perfectly" (p,16); Some children mentioned intimacy and privacy issues: "It may interfere when trying to express yourself, you should not have technology such as this when trying to express yourself. Think about a situation where you want to tell a friend a secret and the robot just sits there, interfering, you just can't do it (express yourself)" (p.25).

6 DISCUSSION

Our findings reveal that children as young as 10-13 years old are able to understand basic ML concepts. Furthermore, children were able to apply their understanding to other ML related situations in the real world, and were able to come up with their own accurate and meaningful ideas for ML applications. As children are growing up in an increasingly ML infused world, an accessible, motivating, direct experience

with underlying ML processes will enhance their ability to generate accurate ML mental models.

Children's ability to understand ML processes was evident only when the learning experience involved iterations between the Data Labeling and Evaluation building blocks. Hands-on experience with just one ML building block did not contribute to children's understanding. We can therefore conclude that iterations between labeling and evaluation of data is needed to support learning. The evaluation should provide a real-time feedback on the system's accuracy, indicating if the sampling is sufficient. This type of iterative, hands-on experience involves a reflection process previously shown to be necessary for learning [11, 40]. The Partial System conditions did not result in a smaller learning effect, but in no learning at all. When children were not provided with an opportunity to both label data and receive feedback they could not construct accurate understanding. This implies that black-boxing too many processes can be similar to black-boxing all processes. In the Partial System conditions, as well as in the pretest, most children provided inaccurate explanations for the ML scenarios and did not show any understanding of ML processes. For example, children explained how GPS data can help a system recognize tennis gestures. Such explanations are based on some understanding of technological processes, but are inaccurate. These may be due to inaccurate or partial mental models that children construct when not provided with appropriate experiences to generate understanding, presenting a risk as inaccurate or partial mental models are difficult to overcome [16]. Future research should explore this idea and identify the balance between uncovering carefully selected underlying processes, while providing a "Low Floor" experience for children [32]. In our system, uncovering two building blocks and keeping the others black-boxed, balanced the learning experience. It is possible that more ML building blocks could have been uncovered without interfering with the learning process. Future work should test additional conditions.

In sum, ML learning systems for children should allow for direct experience with sufficient building blocks. Children should be able to collect and classify data by themselves and evaluate their sampling using feedback of the system's accuracy. This process of Data Labeling and Evaluation evokes reflection and iterations that are essential for learning. We further point out that the learning system we designed involved structured learning tasks as well as noisy data generated by the physical input device, both providing opportunities for self-learning through debugging and overcoming challenges. The combination of a system that uncovers specific black-boxes and a learning experience which involve appropriate challenges, led children to perform cycles of trial-and-error in an effort to improve recognition accuracy, which contributed to learning and understanding.

7 LIMITATIONS

The present study has several limitations. We chose to focus on supervised ML and specifically on classification tasks, considered to be the simplest form of ML [1], future work should study if children are able to understand more complex ML processes . The learning process was based on a hands-on experience with a ML system, and was not compared to non hands-on learning, future work should evaluate if children are able to understand ML processes through additional learning methods. Participants' gender wasn't balanced (20 boys and 10 girls). We made our best effort to recruit a balanced sample, but faced a gender imbalance in Scratch related activities, which was reflected in our participants. Taking this limitation into account, we made sure participants were balanced between the conditions. Furthermore, we tested and found there were no gender effects on understanding.

8 CONCLUSION

In this work we showed children are able to understand ML processes, apply their knowledge in different contexts, and generate accurate and meaningful new examples for real-world ML applications. Furthermore, we revealed that a direct experience with Data Labeling and Evaluation leads to iterations from assumptions to feedback, which in turn leads to understanding. We believe this process is instrumental to the formation of accurate mental models.

Our recommendation for ML product designers is to carefully consider the impact of their work on children. Direct experiences with accessible ML building blocks should be integrated into products, uncovering black boxes and allowing children to collect data by themselves, receive feedback on system's accuracy, and iterate in a process of trial and error. In addition, we encourage educators to provide children with more opportunities for direct experience with ML building blocks in an iterative process of trial-and-error.

ACKNOWLEDGMENTS

We would like to thanks Denis Triman, Yoav Wald, Nadav Viduchinsky, and Adam Agassi for their help in different stages of this study. This research was supported by the Scratch Foundation.

REFERENCES

[1] Ethem Alpaydin. 2009. *Introduction to machine learning.* MIT press.
[2] Alissa N Antle. 2017. The ethics of doing research with vulnerable populations. *interactions* 24, 6 (2017), 74–77.
[3] Brad Astbury and Frans L Leeuw. 2010. Unpacking black boxes: mechanisms and theory building in evaluation. *American journal of evaluation* 31, 3 (2010), 363–381.
[4] Sharon Bertsch, Bryan J Pesta, Richard Wiscott, and Michael A McDaniel. 2007. The generation effect: A meta-analytic review. *Memory & cognition* 35, 2 (2007), 201–210.

[5] Christopher Bishop. 2006. Pattern Recognition and Machine Learning. *Pattern Recognition and Machine Learning* (2006).

[6] OL Davis Jr. 1959. Children Can Learn Complex Concepts. *Educational Leadership* 17, 3 (1959), 170–75.

[7] John Dewey. 1998. *Experience and education.* Kappa Delta Pi.

[8] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. Hey Google is it OK if I eat you?: Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children*. ACM, 595–600.

[9] Stefania Druga, Randi Williams, Hae Won Park, and Cynthia Breazeal. 2018. How smart are the smart toys?: children and parents' agent interaction and intelligence attribution. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 231–240.

[10] Nicholas Gillian and Joseph A Paradiso. 2014. The gesture recognition toolkit. *The Journal of Machine Learning Research* 15, 1 (2014), 3483–3487.

[11] George Hein. 1991. Constructivist learning theory. *Institute for Inquiry. Available at:/http://www. exploratorium. edu/ifi/resources/constructivistlearning. htmlS* (1991).

[12] Elliot Hirshman and Robert A Bjork. 1988. The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14, 3 (1988), 484.

[13] Tom Hitron, Idan David, Netta Ofer, Andrey Grishko, Iddo Yehoshua Wald, Hadas Erel, and Oren Zuckerman. 2018. Digital Outdoor Play: Benefits and Risks from an Interaction Design Perspective. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 284.

[14] Tom Hitron, Iddo Wald, Hadas Erel, and Oren Zuckerman. 2018. Introducing children to machine learning concepts through hands-on experience. In *Proceedings of the 17th ACM Conference on Interaction Design and Children*. ACM, 563–568.

[15] Cindy E Hmelo and Mark Guzdial. 1996. Of black and glass boxes: Scaffolding for doing and learning. In *Proceedings of the 1996 international conference on Learning sciences*. International Society of the Learning Sciences, 128–134.

[16] Cindy E Hmelo, Douglas L Holton, and Janet L Kolodner. 2000. Designing to learn about complex systems. *The Journal of the Learning Sciences* 9, 3 (2000), 247–298.

[17] Nina Holstermann, Dietmar Grube, and Susanne Bögeholz. 2010. Hands-on activities and their influence on studentsâĂŹ interest. *Research in Science Education* 40, 5 (2010), 743–757.

[18] David H Jonassen and Philip Henning. 1996. Mental models: Knowledge in the head and knowledge in the world. In *Proceedings of the 1996 international conference on Learning sciences*. International Society of the Learning Sciences, 433–438.

[19] Ken Kahn and Niall Winters. 2017. Child-friendly programming interfaces to AI cloud services. In *European Conference on Technology Enhanced Learning*. Springer, 566–570.

[20] Alison King. 1992. Facilitating elaborative learning through guided student-generated questioning. *Educational psychologist* 27, 1 (1992), 111–126.

[21] Janet L Kolodner, David Crismond, Jackie Gray, Jennifer Holbrook, and Sadhana Puntambekar. 1998. Learning by design from theory to practice. In *Proceedings of the international conference of the learning sciences*, Vol. 98. 16–22.

[22] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. 2007. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering* 160 (2007), 3–24.

[23] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. ACM, 126–137.

[24] Silvia Lovato and Anne Marie Piper. 2015. Siri, is this you?: Understanding young children's interactions with voice input systems. In *Proceedings of the 14th International Conference on Interaction Design and Children*. ACM, 335–338.

[25] Carl E McFarland, Trudy J Frey, and Deborah D Rhodes. 1980. Retrieval of internally versus externally generated words in episodic memory. *Journal of Memory and Language* 19, 2 (1980), 210.

[26] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. 2016. Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research* 17, 1 (2016), 1235–1241.

[27] David E Penner, Nancy D Giles, Richard Lehrer, and Leona Schauble. 1997. Building functional models: Designing an elbow. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 34, 2 (1997), 125–143.

[28] Jean Piaget. 1973. To understand is to invent: The future of education. (1973).

[29] DD Poudel, LM Vincent, C Anzalone, J Huner, D Wollard, T Clement, A DeRamus, and G Blakewood. 2005. Hands-on activities and challenge tests in agricultural and environmental education. *The Journal of Environmental Education* 36, 4 (2005), 10–22.

[30] Janet Read. 2015. Children as participants in design and evaluation. *interactions* 22, 2 (2015), 64–66.

[31] Mitchel Resnick, Robbie Berg, and Michael Eisenberg. 2000. Beyond black boxes: Bringing transparency and aesthetics back to scientific investigation. *The Journal of the Learning Sciences* 9, 1 (2000), 7–30.

[32] Mitchel Resnick and Brian Silverman. 2005. Some reflections on designing construction kits for kids. In *Proceedings of the 2005 conference on Interaction design and children*. ACM, 117–122.

[33] Royal Society. 2017. *Machine learning:the power and promise of computers that learn by example.* Technical Report.

[34] Eric Schweikardt and Mark D Gross. 2006. roBlocks: a robotic construction kit for mathematics and science education. In *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, 72–75.

[35] Norman J Slamecka and Peter Graf. 1978. The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory* 4, 6 (1978), 592.

[36] Carol Strohecker. 1999. Construction kits as learning environments. In *Multimedia Computing and Systems, 1999. IEEE International Conference on*, Vol. 2. IEEE, 1030–1031.

[37] Florence R Sullivan. 2008. Robotics and science literacy: Thinking skills, science process skills and systems understanding. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching* 45, 3 (2008), 373–394.

[38] Yunjia Sun. 2016. *Novice-Centric Visualizations for Machine Learning*. Master's thesis. University of Waterloo.

[39] Emel Ültanir. 2012. An Epistemologic Glance at the Constructivist Approach: Constructivist Learning in Dewey, Piaget, and Montessori. (2012).

[40] Norbert Wiener. 1988. *The human use of human beings: Cybernetics and society.* Number 320. Perseus Books Group.

[41] Julia Woodward, Zari McFadden, Nicole Shiver, Amir Ben-hayon, Jason C Yip, and Lisa Anthony. 2018. Using Co-Design to Examine How Children Conceptualize Intelligent Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 575.

[42] Oren Zuckerman, Saeed Arida, and Mitchel Resnick. 2005. Extending tangible interfaces for education: digital montessori-inspired manipulatives. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 859–868.

[43] Oren Zuckerman, Tina Grotzer, and Kelly Leahy. 2006. Flow blocks as a conceptual bridge between understanding the structure and behavior of a complex causal system. In *Proceedings of the 7th international conference on Learning sciences*. International Society of the Learning Sciences, 880–886.