From novices to co-pilots: Fixing the limits on scientific knowledge production by accessing or building expertise

Vineet Pandey
Computer Science
Harvard University
Cambridge, MA
vineet@seas.harvard.edu

Krzysztof Z. Gajos
Computer Science
Harvard University
Cambridge, MA
kgajos@eecs.harvard.edu

Anoopum S. Gupta
Department of Neurology
Massachusetts General Hospital,
Harvard Medical School
Boston, MA
agupta@mgh.harvard.edu

ABSTRACT

The current approach of relying primarily on institutional experts to create knowledge to solve humanity's problems is insufficient to meet the scale, diversity, and novelty of people's needs. Building expertise in people to create knowledge they need provides a promising approach. Despite having contextual insights, people fail to rapidly generate sound plans—like experiments—and correctly implement specific actions—like data acquisition and analysis. The limits to progress in multiple domains—like science and healthcare—can potentially be expanded by building procedural expertise among motivated non-experts so that they can build on their contextual insights to create valid and generalizable knowledge. In this paper, we report on the design and evaluation of tools that highlight two ways to realize this vision. First, Hevelius is a motor impairment assessment tool for patients to conduct neurological assessments online. A rare disease community has provided fine-granular data and insights from their homes that current in-clinic assessments fail to capture. Second, Gut Instinct is a social computing system that supports procedural knowledge acquisition for experimentation. A fermentation community used Gut Instinct to successfully design and run between-subjects experiments to test their intuitions. These results suggest exploring ways of producing knowledge that are distinct from the dominant model of institutionally-situated experts testing their ideas on subjects in a lab or a clinic. More constructively, these systems demonstrate how knowledgeable and committed people can be aided and amplified by technology in creating scientific knowledge.

CCS CONCEPTS

• Human-centered computing • Human computer interaction (HCI) • Collaborative and social computing systems and tools

KEYWORDS

Social computing; Sociotechnical systems; Citizen science

ACM Reference format:

Vineet Pandey, Krzysztof Z. Gajos, Anoopum S. Gupta. 2020. From novices to co-pilots: Fixing the limits on scientific knowledge production by accessing or building expertise. In *7th International Conference on ICT for Sustainability (ICT4S2020), June 21–26, 2020, Bristol, United Kingdom*. ACM, New York, NY, USA, 9 pages.

https://doi.org/10.1145/3401335.3401814

ICT4S2020, June 21–26, 2020, Bristol, United Kingdom

ACM ISBN 978-1-4503-7595-5/20/06 $15.00
https://doi.org/10.1145/3401335.3401814

1 INTRODUCTION: LIMITS TO EXPERT TOOLS CONSTRAIN KNOWLEDGE PRODUCTION

Our societies lack enough experts; for instance, most countries have a severe shortfall of teachers, doctors, and researchers [34,38]. At the same time, many citizens possess contextual knowledge to make useful contributions. For instance, Dana Lewis—a Type-1 Diabetes patient with no professional training in medical devices—created her own device to automate insulin delivery to keep blood glucose in a target range. She described her experience and plans in an empirical paper at the American Diabetes Association [11]. Institutional experts have subsequently collaborated with her. This example suggests that people do not need specific institutional credentials to contribute to science. Could this be true for scientific knowledge production in general? We do not know; most citizen contributions are limited to providing data as research subjects. The lack of greater citizen partnership in scientific work is not just an academic concern; people's involvement in science shapes broader public trust in scientific knowledge and expertise [4]. In this paper, we argue for a future of collaborative knowledge production between communities and experts; communities' motivation and resourcefulness provide a starting point.

People's lived experiences provide them contextual expertise at their tasks [16]. Consider bakers trying to make better bread by trial and error or patient communities with movement disorders trying out different devices to improve their gait. Such folks are motivated to improve their situations and possess vital contextual

knowledge needed for success with real-world problems. Some of their lessons could even provide potentially novel and generalizable knowledge. Bakers could inform the science of yeast and patients could provide reports that inform the future design of tools for mobility. Furthermore, many communities already share knowledge among themselves via online fora conversations (patientslikeme.com), product reviews, and blogposts. However, many such insights stay beyond the realm of institutional experts like scientists. While public contributions can create generalizable scientific knowledge, even motivated people lack the knowhow and the tools to do so. We find this to be missed opportunity.

The main contribution of the paper is supporting communities in creating valid and generalizable scientific knowledge. In this paper, we expand ideas of Limits beyond material and physical limits to limits in knowledge production. To support collective knowledge production, this paper identifies two problems faced by communities: 1) finding ways to work with experts; and 2) developing procedural scientific expertise with little effort. This paper makes the following contributions:

1. Two deployed sociotechnical systems to support expert-community collaboration and identify the social preconditions for the success of these systems
2. Contribution mechanisms and just-in-time procedural guidance that mitigate the lack of prior knowledge
3. Empirical results testing the efficacy of these approaches with two distinct communities

2 From novices to co-pilots: What do we need

In this section, we discuss the differences in how experts and citizens use online platforms to create scientific knowledge, and provide ways to deepen contributions by citizens and communities.

2.1 Experts possess the intent and capacity to use new technologies

Institutional science has benefitted immensely from large-scale global collaboration. Possessing both the intent and capacity to contribute using the internet, experts in many fields have radically changed how they perform science. Experts benefit from conceptual knowledge, professional training, pre-existing organizational structure for collaboration, and direct access to resources. For instance, LIGO's pathbreaking discovery of gravitational waves brought together over 100 researchers from over 100 institutions across 18 countries (ligo.org/about). Scientists increasingly share data and results faster (arxiv.org). Large scientific projects, like the Human Genome Project, took to agile science by sharing methods, data, and insights to collaboratively speed discoveries. Scientists also form global collaborations to accelerate research in nascent scientific domains, like the Earth Microbiome project (earthmicrobiome.org). Efforts to further expand participation in scientific research are bearing fruit: *Lab in the Wild* recruits anyone with an internet connection for behavioral studies [29]; *All of Us* aims to recruit one million Americans from all strata of society (allofus.nih.gov). Distributed data contributions from people around the world—browsing online [10], using activity trackers, and joining scientific projects—have enabled valuable insights on topics including obesity [2], aesthetic preferences [30], sleep [13], and the human microbiome [23].

2.2 Communities contribute to science when their intent is supported by experts or tools

When citizens participate in science, it is typically as embedded sensors that are aggregated by experts. Public involvement in scientific endeavors continues to be largely limited to performing tasks just beyond the reach of computers. A classic example is Audubon's Christmas bird count, run since 1900 [3]. Online examples include reporting flower blooms in Project Budburst [5]; and identifying galaxies from satellite imagery in GalaxyZoo [39]. To support motivated communities in performing more complex scientific work, we see two approaches: collaborating with experts, and developing task-specific expertise.

Collaborating with experts

Patient communities are intrinsically motivated to expand on existing knowledge for their medical conditions. Many communities use online fora to share caregiving information and discuss research progress. However, contributing to scientific research requires greater commitment and knowhow. Collaborating with experts provides one way for communities to contribute. Experts' knowledge and skills provide the confidence that community efforts would not be wasted. Amyotrophic lateral sclerosis (ALS) patients designed and ran a study to test the efficacy of lithium in reducing their symptoms; this effort was led by the PatientsLikeMe platform creators with significant experience in study design [35]. The results of this study foreshadowed what a NIH-funded study found months later. While the PatientsLikeMe experiment provided specific ways for ALS community members to contribute, this need not hold for other conditions. Rare diseases provide an example.

Rare diseases are disorders that affect fewer than 200,000 people. This quantitative distinction in the number of patients leads to differences in the availability of experts (both in numbers and location), quality of care, general awareness about the condition, and current state of research [17]. Rare diseases provide an extreme example where patients' inputs can potentially create much-needed knowledge; however, accessing enough patients with these conditions is difficult. For example, University of Utah Computer Science Professor Matthew Might contributed his own resources, tapped his intellectual network, and used his scientific prowess to better understand a loved one's rare disease [24]. In the absence of systematic ways of finding and working with experts, it falls on individuals to put in exceptional efforts and resources to perform complex tasks and to reach out to experts. We need a participatory approach where experts and citizens help each other answer questions.

Expert-community collaboration—such as the PatientsLikeMe experiment—requires both social and technical successes. Sharing objectives, building trust, and respecting constraints are critical to getting started. Both the community and the expert leaders need to spend their social capital to ensure that the collaborative effort meets all the stakeholders' objectives. Experts might need clinically meaningful data while community members might be interested in answering their own questions. Bridging the gap between the two might require multiple rounds of conversations and iterations to come up with a study design that is not just scientifically correct but also community validated. People must also trust each other to perform the appropriate steps. Perhaps unsurprisingly, the high-risk Lithium experiment on PatientsLikeMe was led by the platform creators who had gained the trust of the community over multiple years. Finally, the tools developed must also respect the constraints of the community and the experts. For instance, both patients and experts might be strapped for time, so making data collection least burdensome is important. Such real-world studies must also provide mechanisms for community members to voice their questions or drop out without any fear.

Software to develop task-specific expertise in people

While expert-community collaboration provides one way to create scientific knowledge, it might not always be suitable. Experts are in short supply and communities and expert objectives might not always align. In the absence of experts, translating motivation to action requires developing relevant capacity; this is especially true for complex domains. Making a challenge visually salient is an effective way to on-board novices. One such domain is biochemistry games: finding protein structures in Foldit [8], synthesizing RNA molecules in EteRNA [20], and aligning nucleotide sequences in Phylo [18]. Foldit introduced 3D game for specifying low-energy protein structures via direct manipulation [8]. At their best, these citizen science platforms yield novel insights. For example, Foldit players discovered protein structures that helped scientists understand how the AIDS virus reproduces [9]. For tasks that do not have as a crisp visual analogue as protein folding, people need better support for learning because they lack the years of domain training**.**

To create knowledge, they need mental scaffolds for organizing complex work, domain knowledge to compose and execute the steps, and ways to ask for help. Conceptual learning—the primary focus of classroom teaching—involves understanding and interpreting concepts and the relations between concepts. In contrast, procedural learning teaches "action sequences for solving problems" [31]. Success with complex creative activities requires procedural knowledge (how to do things) in addition to conceptual knowledge (facts). While many resources offer facts, procedural learning is often ignored. To contribute usefully, people need to have a good working model of both the concepts and procedures for an activity.

Even with learning resources, complex tasks can be unwieldy to manage. Complex tasks can be made manageable by dividing them into distinct phases. Touchstone demonstrates the power of a semi-automated workflow integrating experiment design, testing, and analysis [21]. Crowdsourcing has similarly innovated by creating distinct phases: break larger tasks into microtasks; algorithms specify the division, dependency, and agglomeration activities while workers perform small tasks supported by task-specific guidelines [19].

Our research builds on prior work in the Limits community at both conceptual and systems levels. Barath and Pargman discuss the importance of refactoring society's complex challenges to reduce complexity and involve more people in seamless ways [28]. Our work provides such refactoring for experimentation and neurological assessment. Penzenstadler et al. described techniques that provide guidance to people building a smart garden [27]. We use related ideas of procedural guidance to support knowledge production. In this paper, we report on the design and evaluation of tools that demonstrate ways in which communities can collaborate with experts and build expertise. These systems demonstrate how knowledgeable and committed people can be aided and amplified by technology in creating scientific knowledge.

3 Community-expert collaboration to understand a rare disease

Ataxia-Telangiectasia (A-T) is a rare inherited neurological disorder. Typically apparent during childhood, this disorder is characterized by impaired coordination of movement, impaired immunity, increased cancer risk, and telangiectasias (small widened blood vessels). Since it affects multiple body systems, A-T requires a complex care team comprising multiple specialists making it extra daunting for caregivers to both understand and manage the condition.

3.1 Hevelius for remote neurological assessments

Hevelius is an online motor impairment assessment tool for patients with neurological conditions. Hevelius has the potential to give experts access to larger quantity and more frequent data to better characterize a neurological condition. The patient community can contribute to improving scientific knowledge about the condition while also potentially answering their own questions. Hevelius is not too dissimilar from prior attempts to "crowdsource" the collection of scientifically relevant data. While helpful, we will argue next that its success critically depends on a set of social preconditions that may be difficult to satisfy for many disease communities.

Building trust and identifying objectives take precedence

Building trust and agreeing on research objectives is critical for community and experts to collaborate. This requires open communication and at least partly shared mental models about the topic---the rare disease. Such trust-building and knowledge sharing can be mediated and accelerated by trusted organizations

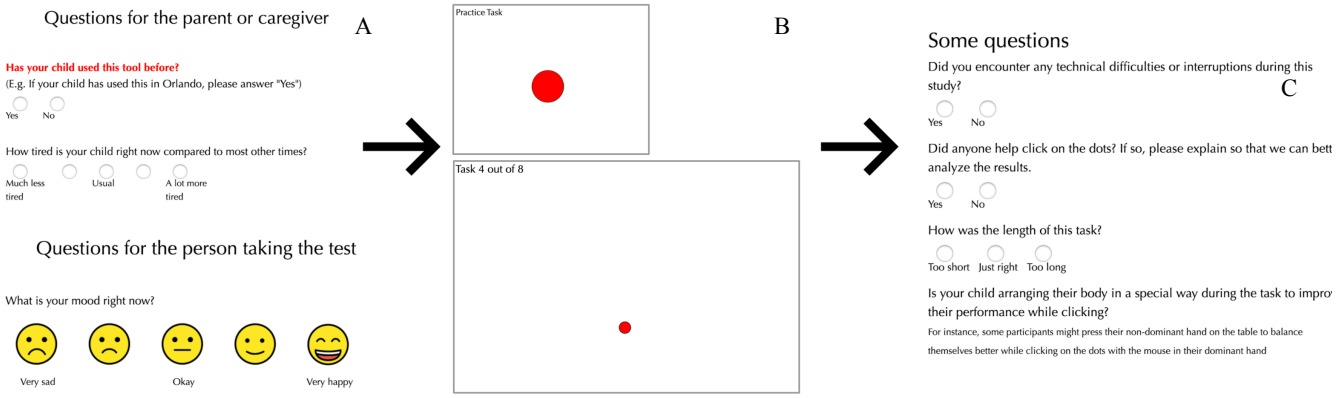

Figure 1: With Hevelius, rare disease community members provide researchers weekly well-being self-reports, motor performance data, and insights about tool usage. The userflow for Hevelius: A) Caregivers and participants answer questions about lifestyle and well-being; B) Participants perform practice tasks to warm up and then click on eight rounds of nine dot clicks; C) Caregivers answer questions about the participants' experience using the tool.

that 1) understand experts' and community objectives, and 2) communicate expectations from both parties to collaboratively shape the project. Here we describe the social processes that happened for this project to come to fruition.

One member of the research team is a practicing clinician for the rare disease. Apart from possessing research expertise about the condition, his clinical expertise also brings a deeper understanding of patient needs and challenges. Building on this initial trust, it is important to understand the broader rare disease community's objectives. We have received immense support in this step from the relevant rare disease foundation: Ataxia-Telangiectasia Children's Project. With their experience working with the community and multiple experts, Ataxia-Telangiectasia Children's Project provides multiple contributions that are difficult to achieve for a small research group. First, they provide a consistent point of contact to reach out to the community members who already trust them. Second, they have accelerated the research cycle by sharing insights on potential plans based on their prior experience supporting expert-community research. Third, they have actively sought to share resources among multiple experts who could potentially collaborate on similar topics. In short, Ataxia-Telangiectasia Children's Project provides both access to the community and to other experts; developing these relationships ourselves would require substantial effort.

Regarding community objectives, prior work has demonstrated that families have complex knowledge, caregiving, and emotional needs [17]. Planning to meet some of these objectives both improves the research and makes the experience more rewarding for the community. Many community members shared their needs of better understanding their loved ones' condition and progression more objectively than their daily observations. Based on our conversations with them (described later), we realized that providing near-term benefits by showing relevant data back to the participants in the next tool iteration is both doable and potentially valuable for the community. Given this convergence in objectives, remote data collection with a clinically validated tool provided one concrete way to proceed.

Tool Implementation

Hevelius comprises a computer mouse-based tool that provides objective, granular, interpretable, multidimensional quantification of motor impairment in the dominant arm with just a few minutes of use by the patient [14]. Compared to a standard neurological exam and existing disease rating scales, Hevelius does not require expert judgement to compute the scores and it provides assessments that are more granular. The data collected by Hevelius has been used to accurately measures disease severity, to distinguish between different neurological disorders like ataxia and parkinsonism, and to captures disease progression.

In addition to the motor impairment measurement component, Hevelius also includes questionnaires for collecting: 1) self-reports on health and lifestyle; and 2) self-reports about using the tool (Figure 1). Throughout this section, we refer to a user with neurological disorder as a participant and their family members overseeing their tool use as caregivers. We refer to a unique {participant, caregiver(s)} set as a family.

Health and lifestyle self-reports ask the family about well-being of the participant and significant events since last use of the tool. Finally, at the end of tool usage, self-reports about using the tool intend to get at the family's contextual insights integrating their experience with the tool and their observations. *E.g.*, one question asks families about how participants might be trying to improve their performance by altering their body and arm posture.

3.2 Research questions

Hevelius was first used for two years in a movement disorders clinic, primarily with ataxia and parkinsonism patients. In that context, only the motor impairment assessment part of the tool was presented to the patients and the use of the tool was supervised by a trained technician. The data collected by Hevelius were later compared to results of same-day traditional neurological assessments and were shown to be predictive of the diagnosis and disease severity [14].

Building on Hevelius' success in the in-clinic deployment, we next collaborated with a particular rare disease community to answer two questions: First, does the tool provide reliable data across multiple uses (including in-clinic visit)? Second, what challenges do people face in using this tool at home over multiple weeks? Furthermore, we also wanted to know whether families' contextual insights and feedback help experts develop novel ideas. To answer these questions, we ran a 10-week study where people were requested to use the tool at home once a week.

3.3 Study Design

This study has three components: 1) in-clinic visit, 2) at home deployment, and 3) final interview (pending).

In-clinic visit: Caregivers and participants used the tool under the observation of the primary author. The in-clinic visit was done at an annual rare disease meetup in January 2020. Families traveled to the meetup for regular clinical assessment and social events; they did not travel to just use the tool. Families were requested to bring images of their setup at home to receive feedback on how to collect best data (*e.g.*, by reducing distraction near the setup). This visit provided three benefits: 1) baseline data for the research team to compare at home usage data to, 2) clarifications and corrective feedback for the families from the researchers, and 3) calibration of the dot size for individual participants; since participants' severity of the condition varies from mild to severe, using the same dot size was infeasible: it would be too easy for some and too difficult for others. Additionally, one goal of meeting families was to build trust between the research team and the participants by engaging over the task as well as via conversations in an informal setting [32]. Families were provided identical mice to take home; they were asked to note down a day and time of the week for using the tool once a week starting two weeks after settling at home.

At home deployment: Participants were requested to use the tool once a week; they received email updates from the community coordinator of Ataxia-Telangiectasia Children's Project, and communicated with the research team members via emails and text messages if they faced a concern with the tool. Typically, if a person did not use the tool for two weeks, the coordinator would reach out and enquire if they had any issues. The research team did a weekly debrief to look through the results, identify outliers, see usage data, and tweak the tool.

Final interview (pending): The study is ongoing and we expect to complete the interviews in June.

3.4 Results

21 families used the tool at home. 18 started with in the in-clinic visit in January 2020; 3 joined later. Families primarily used Hevelius on weekends.

Does the tool provide data that is reliable across multiple use including in-clinic visit? Yes. Out of the 32 features tracked by Hevelius, researchers found the ones that are reliable across weekly usage. (Figure 2A)

Did people continue using the tool over multiple weeks? Yes. (Figure 2B).

3.5 Discussion

Why and how did the community use the tool?

Why did people choose to participate in the at home deployment? From our in-clinic conversations with caregivers, three reasons emerged: helping the community; answering their own concerns; and altruism. Many participants mentioned that the community's

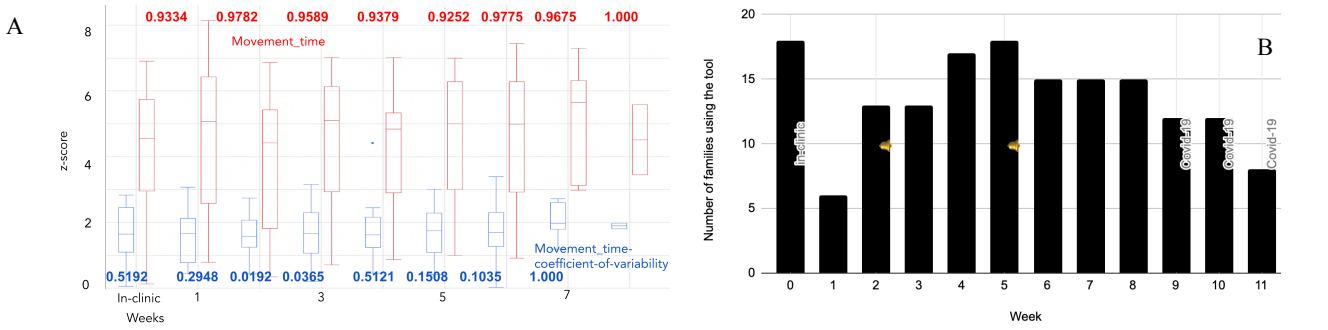

Figure 2: Results: A) The values in decimal show the correlation between usage during $Week_{n+1}$ with $Week_n$. Examples of measures with high and low reliability: movement time (in red) of the pointer was highly correlated across the weeks; movement time coefficient of variation (in blue) was not. B) Tool usage varied (🔔= reminder sent; Covid-19=Stay at home orders)

support had helped them over the years; such support included receiving details about doctors, resources about understanding the condition, lifehacks for common concerns, and emotional support. Many felt that they were doing their part in helping the community by participating in this study. People also shared questions that they felt Hevelius data could help answer; *e.g.* some families wanted to check whether an experimental treatment (that the participant was enrolled in) was working; some wanted to use this data to check and refine their understanding observing the participants' behavior. Finally, many participants also mentioned that they just wanted to help out. They felt that there is little known about the condition and their contributions could potentially create new knowledge that might be useful down the road.

Did people use the tool "properly"? While it is not possible to exactly know how families used the tool remotely, the quality of data being comparable to the in-clinic visit provides some evidence that people persisted with using the tool as guided. This is important: receiving meaningful data for clinical assessment from home implies that this tool could potentially be used with other communities. Why did people persist in using the tool? We believe our final interview will provide us insights about this question. From both who persisted in their usage and those who did not, we intend to learn more about their motivation, the burden, fatigue, and utility (perceived/real) from continued usage.

Experts can rapidly test hypotheses and improve the tool with task data and caregiver responses

Switching tasks to test ideas: Running the study remotely on an online platform enables the clinicians to switch the task without much effort. For instance, researchers wanted to compare two types of clicking tasks: one where the dots show up one at a time in a "random" (to the user) location and another where the dot shows up at diametrically opposite locations around the center of a circle. Early analysis verified that participants took less time on the second task; this can be potentially useful for those with severe impairments who struggle more on the first task.

Learning about different strategies: During the in-clinic visit, many participants demonstrated strategies to better use their dominant hand while clicking. Caregivers reported that this was true at home: many participants used their non-dominant hand to balance themselves on the desk while clicking with their dominant hand. Some participants would move closer to the screen when clicking on smaller dots. Such contextual insights help researchers identify how participants might tailor their behavior while using the tool and provide potential confounding factors.

Providing support to manage a diversity of setups: Browsers provide similar "platforms" to maintain consistency. While diversity in browsers can be problematic at times, we did not hear about any major issues. Some families had old browser versions; we worked with them remotely to upgrade their browser by sharing instructions. Others had too low screen resolution to see all the dots on the screen. By tracking people's screen resolution, we edited our tool to make it work with their home setup.

Improving Hevelius to meet more objectives

Provide appropriate feedback to meet community's needs: Some families had earlier mentioned their hope from using the data to test their ideas. We are currently in the process of analyzing and understanding the results thoroughly ourselves before sharing them with the community. There are two reasons: 1) analyzing and understanding the data ourselves first makes it easier to take the communities' questions; 2) some data can be potentially emotionally burdensome, especially if the participants' task data seems to signal a downwards trend. For such situations, we need to carefully present and explain appropriate data elements.

Tackle common slips: The research team provided families with identical mice to 1) maintain a consistent input device across families and weeks; and 2) reduce friction in getting started. Some families reported participants struggling with using the mouse at home; most reported being used to tapping on tablets; they had difficulty holding the mouse or consistently clicking on the left dot. Developing ways for participants to use the mouse more comfortably is a relevant task.

Hevelius demonstrates that communities can work with experts to generate useful data and insights that can potentially meet both groups' needs. However, access to experts might not be easy for everyone and sometimes, not necessary. Gut Instinct demonstrates how communities can self-organize to answer their questions.

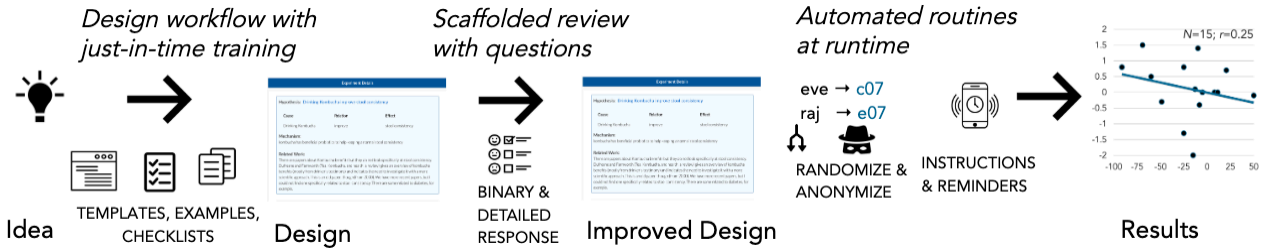

Figure 3: Gut Instinct enables anyone to design and run experiments to test their intuitions by integrating conceptual learning embedded via short lectures and software-guided procedural learning. Experiment creators can invite anyone to review and participate in the experiment. Participants from around the world join experiments, follow instructions, and provide data in response to automated data collection reminders.

4 Community-run experiments to understand the human microbiome

The human microbiome is the collection of all microbes and their genetic components in and on our bodies. As a scientific domain, it is nascent, highly contextual (people's microbiome is unique), and personally motivating (altering the microbiome can improve health). Each of us hosts a different collection of microbes, and this collection is influenced by our environment, diet, health, lifestyle, and genetics. A major scientific effort is to better characterize and understand this diversity and the causal factors for it (hmpdacc.org). This requires engaging diverse participants at scale. How can people's situated knowledge supplement institutional science?

4.1 Gut Instinct: From intuitions to experiments

Gut Instinct (Figure 3) is a social computing system that provides non-experts with just-in-time procedural knowledge necessary for designing and conducting rigorous scientific experiments. Importantly, Gut Instinct provides mechanisms to support productive collaboration during all steps of the experiment design, data collection and analysis among a larger community of non-expert enthusiasts. The system enables people people in designing experiments, getting them reviewed, and running them with interested participants. GI enables knowledge acquisition in two ways: 1) reifying conceptual bits in the software; and 2) providing procedural guidance with examples, checklists, and templates. Gut Instinct is a general platform but its initial design as geared towards improving our understanding of the microbiome.

What makes experimentation difficult? Despite a predetermined goal and a formalized process, experimentation requires making contextually-appropriate decisions [22]. Good experiment design is inherently user centered; designers need awareness of others' interpretation of their ideas and asks. Providing feedback on experiment designs requires knowing the success criteria and how to help improve. Finally, successfully running an experiment requires managing multiple processes such as random assignment, anonymizing participant details, and sending instructions and reminders for data collection.

4.2 Design-Review-Run: From Intuitions to Investigations

Gut Instinct creates different roles and supports them with procedural guidance. Role-based approaches confer three benefits: 1) clean delineation of responsibilities improves chances of task completion, 2) clustering similar tasks reduces overhead and increases consistency; 3) people can decide their contribution levels. Gut Instinct requires three roles for each experiment: designer, reviewer, and participant. Gut Instinct offers procedural support for each: 1) a design workflow provides just-in-time training, 2) review with scaffolded questions, and 3) automated routines for runtime activities like data collection. Users form and refine with the help of contextual support and learning resources from the system.

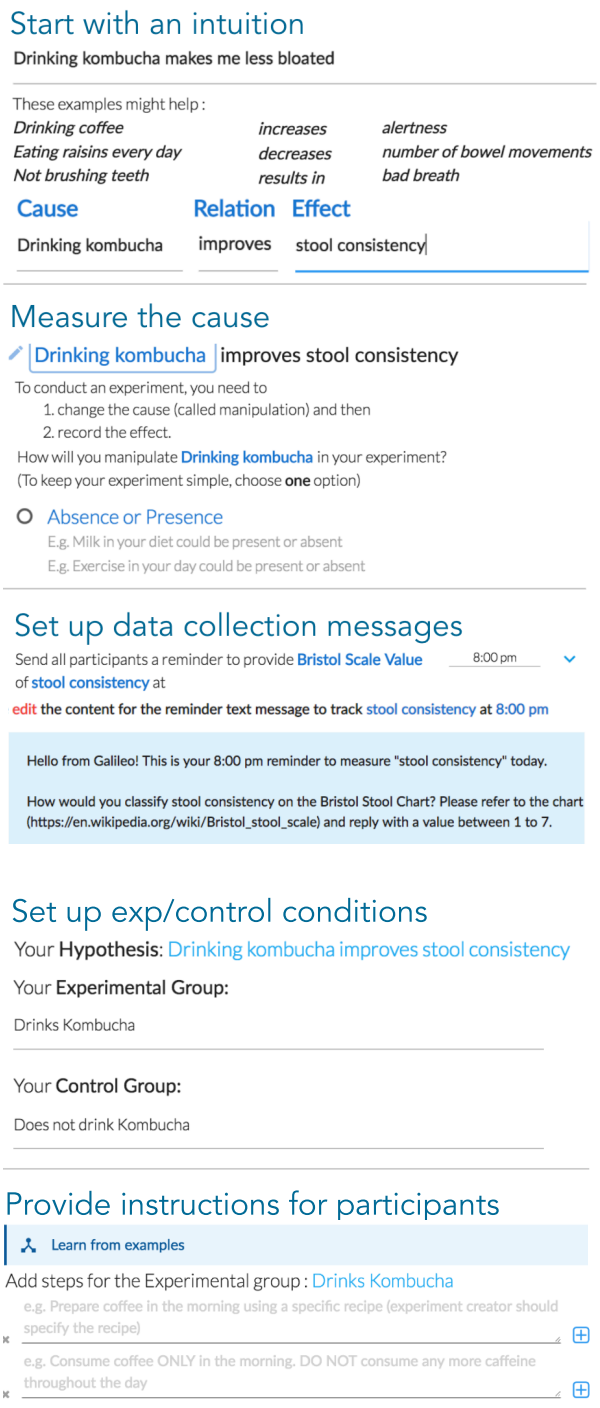

Figure 4: Gut Instinct's design module helps people transform intuitions into experiment designs. It walks people through 1) converting an intuition to a hypothesis, 2,3) providing ways to manipulate/measure cause and effect, 4-5) specifying control and experimental conditions, and (not shown) providing inclusion/exclusion criteria.

Design an Experiment from an Intuition

People have many, often poorly-framed, hypotheses. Gut Instinct's design workflow helps people harvest and sharpen them (Figure 4). Examples illustrate possible choices and how they relate; templates provide structure; and embedded videos explicate technical issues. Such procedural support can improve on-task performance [26]. A final self-review step provides an overview of the experiment. The design workflow does not mandate double-blindness or the use of placebo; designers can choose to specify these details.

Review the Design via Feedback from Others

Gut Instinct requires at least two reviews before an experiment can be run. The designer invites reviewers: an online community member, a teacher, or anyone else who can provide useful feedback. Upon receiving reviews, the designer edits their experiment to address any issues. For research purposes, Gut Instinct logs version changes. Reviewers provide both binary assessment and written responses to specific questions (Figure 5). These questions cover structure (*e.g.*, accounting for confounds), pragmatics (*e.g.*, measuring the real-world cause/effect), and participant experience (*e.g.*, data reminder time). Reviewers are ineligible to be participants in the same experiment. Similarly, creators may not review their own experiment.

Run an Experiment using Procedural Support

To launch an experiment, its designer shares a unique URL with potential participants. Gut Instinct automatically manages four activities to reduce bias and workload:

1. Randomized placement of people into conditions [22].
2. Maintain a per-experiment participant map ([usernames] → [exp_id]) for anonymity
3. Collect and clean data (sending data collection messages and reminders at time-zone appropriate times, parsing the responses, updating participant and experimenter views).
4. Prompt experimenters to perform tasks when conditions are met (*e.g.*, setting the start date when enough participants have joined or reminding participants with missing data).

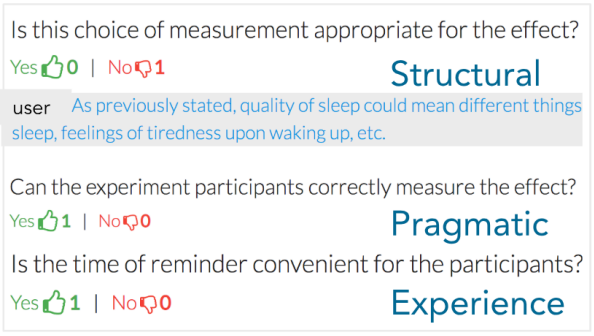

Figure 5: Reviewers walk through an experiment providing binary rubric assessments. A No response prompts reviewers to provide concerns and suggestions.

Participation comprises following instructions (*e.g.*, drink kombucha) and providing self-report responses to platform queries (Figure 6). Self-reports provide the primary data collection mechanism. Participants can optionally answer follow-up questions that capture contextual insights (*e.g.* changes in daily lifestyle due to travel). Gut Instinct presents participant data to

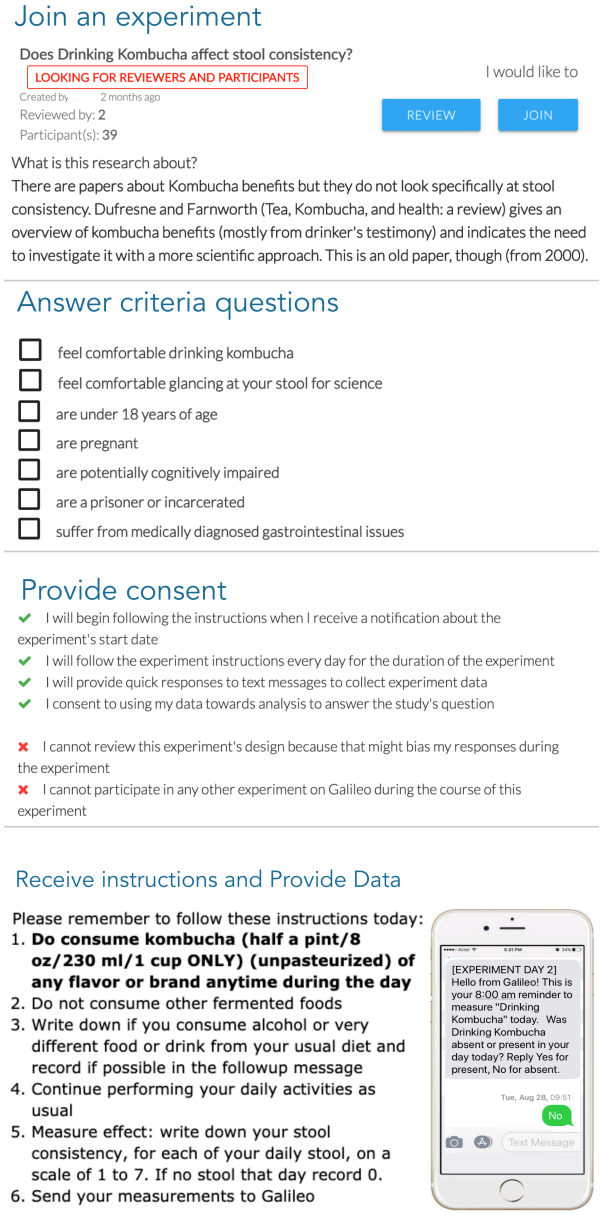

Figure 6: 1) Participants can view a list of experiments. When they elect to join one, they 2) answer inclusion/exclusion criteria, 3) consent to following the provided steps, and 4) receive instructions. Participants receive daily, condition-specific requests, and respond with data and/or clarifying questions.

experimenters using participant ID rather than real name or username. When an experiment ends, Gut Instinct sends a summary of results to participants. Participants can anonymously discuss experiments at the end, so the experimenter and other users on the platform can learn from their feedback. The experimenter's dashboard provides a summary of their experiment's progress and supports lightweight tasks to improve the quality of data collected. The dashboard lists tasks: answer clarifying questions, remind/thank participants, or look at trends in data. Experiments have a minimum participation count; there is no upper limit to the number of participants. People who sign up after a cohort begins are added to a waitlist.

The Gut Instinct web application uses the Meteor (meteor.com) framework for synchronization, Jade for the front end (jade-lang.com), and Materialize for styling (materializecss.com). The current Gut Instinct implementation supports email, SMS with text message gateway Twilio (twilio.com), and WhatsApp. Gut Instinct logs responses to a MongoDB database.

4.3 Study: Fermented foods community designs, reviews, & run experiments

Does drinking Kombucha improve stool consistency? Kombucha is a fermented tea drink popular in many parts of the world. Fermented foods (miso, yogurt, ayran, kefir) have been a staple in many cultures for thousands of years [7]. While there is widespread belief that kombucha "benefits the gut"[1], there is little published empirical evidence for these claims [12]. The experimenter hypothesized that kombucha supplies beneficial probiotics that help maintain normal stool consistency, and designed a between-subjects experiment.

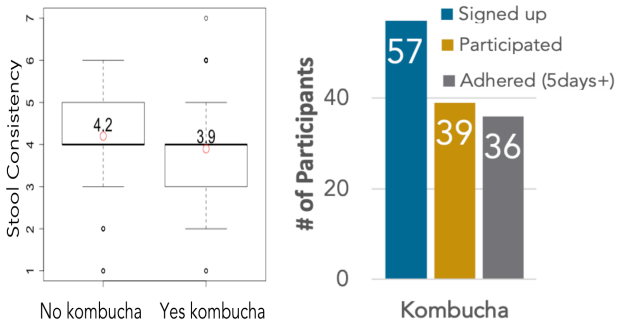

Figure 7: Kombucha community designed and ran an experiment which ran for a week. They found that drinking kombucha improves stool consistency (N=36; p<0.03). After signing up, 68% of people participated in Kombucha; 92% of those who participated reported adherence. Reasons for non-adherence included being busy, annual leave, and brewers needing to check on the taste of Kombucha.

[1] https://www.nytimes.com/2019/10/16/style/self-care/kombucha-benefits.html

4.4 Results

The community-led experiment found evidence that drinking kombucha improves stool consistency (Figure 7).

4.5 Discussion

Before the Experiment

From initial design to launch — 37 days elapsed. The experiment ran for a week.

Design and Review: The experimenter had not previously designed and run an experiment with people but knew some concepts about experiment design. They have a PhD degree in ecology and are a Brazilian national. The experimenter had lived experience of their experiment's topic but had never scientifically studied it. Reviewers provided a total of 32 boolean answers and 12 detailed comments. Comments focused on two themes. First, reviewers helped make the hypothesis and measures more specific. A reviewer criticized the experiment's 5-point Likert scale for bloatedness as overly vague. In response, the experimenter found and adopted the Bristol stool chart—a picture-based scale that is the industry standard [37]. Second, reviewers suggested improving data quality by instructing participants to skip confounding activities. All issues that reviewers raised were tightly connected to Gut Instinct's review rubric (Figure 5). At the end of review, the experiment design used appropriate measures, provided a minimal-pairs design, tracked confounds, and provided appropriate criteria for participation.

Pilots and finding participants: Two lessons emerged. First, some participants were loath to look at their stool. Since viewing one's stool is necessary, the experimenter added an inclusion criterion enforcing this. Second, some participants reported eating other fermented foods in the process; the experimenter modified the instructions for participants to not consume these. The experimenter publicized the experiment on Instagram, Twitter, and newsletter; they also created a poster, and reached out to enthusiasts in their city in Brazil and an American city. After failing to recruit sufficient participants, the experimenter collaborated with a kombucha fermenter in an American city who knew more kombucha enthusiasts.

During the Experiment

Retention: 57 people signed up for the kombucha experiment; 36 completed it (68%). 78% of dropouts occurred in the first 48 hours. The reasons participants reported for dropping out included lack of interest, holidays, and work travel.

Adherence: The experiment garnered 76% adherence: 86% for days of no kombucha, and 70% when asked to drink kombucha. Some participants disclosed confounds and reasons for non-adherence. For example, drinking alcohol was a reported confound, because it might affect kombucha's impact on the body. Similarly, participants' non-adherence reports included scheduled disruptions like travel and holidays and work responsibilities like brewers needing to check on their kombucha.

Data Collection: Most American participants selected text solicitations (86%); participants elsewhere received email solicitations due to varying regulations around automated text messages (*e.g.*, replying to an automated text message in Brazil or India is infeasible since the source number is masked). 56% of participant responses came within 30 minutes of the solicitation; 21% of responses took more than 90 mins. Participants sparingly responded to follow-up questions. The experiment requested that all participants adhere to the protocol as much as possible without harming their health. Participants could ask the experimenter (via the platform) if confused. Participants' clarifying questions focused on measurements (*e.g.*, measuring stool consistency once during the day or multiple times) and specific lifestyle choices (*e.g.*, consuming probiotics while drinking kombucha?). Participants reported an overall positive experience (Figure 8).

5 Discussion

Our research suggests that building tools is not sufficient to meet the objectives of communities producing scientific knowledge. While the tool stayed the same across different communities' usage, successes came about when the communities found value in using the tools. Sociotechnical systems can amplify the efforts of committed and knowledgeable individuals [33]. However, finding such people, getting them started, and keeping them invested requires complementary efforts. The rare disease foundation—Ataxia-Telangiectasia Children's Project—made it easier for us to find committed community members. The foundation's community coordinator also shared sticky knowledge from prior experience: they shared ways to receive quick feedback from the community and reminded participants when they did not use Hevelius. We believe this support has drastically improved our chances of success. By reducing friction

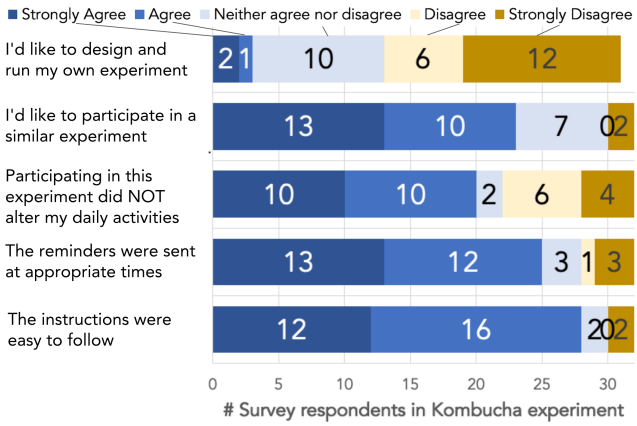

Figure 8: *Kombucha* participants reported an overall positive experience; nearly all expressed an interest in participating in similar experiments (23/32). Most reported that its instructions were easy to follow (28/32) and that reminder times were appropriate (25/32).

at multiple steps, such organizations provide key support in creating and maintaining successful collaborations.

Working with communities requires identifying and respecting their objectives, structure, and existing expertise. The rare disease community was well aware of the importance of their inputs in accelerating research about the condition. Once contacted, community members graciously shared their needs and challenges that have identified future research and development efforts for our work. Furthermore, members cared about the community having helped each other over the years. Such strong social motivation might have been helpful in sustained usage of the tool. For cases where such trust might be more nascent, not visible, or clearly lacking, researchers can better direct efforts towards better understanding the community. Furthermore, the kombucha experiment succeeded when the experimenter reached out to another fermenter with a large online community following via newsletter. Why did people participate once invited? Gut Instinct successfully met fermentation enthusiasts' goals of understanding whether fermented foods such as kombucha improve one's gut health. While this community's objective matched well with the tool, other communities had lesser luck. Another experiment by a lone enthusiast tested the effect of alcohol on time to fall asleep. Despite publicity and initial interest, participants reported low adherence [25]. When there are many individuals interested in a topic but lack a community, techniques like activation thresholds [6] might help make reciprocity explicit. This can also reduce potentially wasted efforts later. We hope to draw from our interviews to better understand participants' fatigue and burden.

While an advanced degree is not a prerequisite for using the systems, having one might confer an advantage. This is unsurprising; contributions to web platforms vary across educational levels. MOOCs are disproportionately completed by learners from more-affluent and better-educated neighborhoods [15], and 73% of citizen scientists and Wikipedia contributors have advanced degrees [1,36]. While all 36 Kombucha participants wanted to participate in future experiments, only two participants wanted to run their own, and both have advanced degrees. While simply asking people to contribute might work for traditional citizen science projects, experimentation might be a bigger leap. This is a humble reminder of how people vary in their intent and/or capacity to use the same tool.

Telemedicine tools can be beneficial in a future of limits. Supporting experts in tracking only specific condition-relevant measures with carefully designed tasks can potentially reduce the amount of data collected, stored, and analyzed. Our database dump runs, for instance, in the low hundreds of MBs. for fifteen participants' eight weeks usage data. Furthermore, such systems can reduce in-person visits to clinicians for patients and their caregivers. Such savings can be substantial for rare disease communities that have to travel far to meet experts, and also for a society increasingly focused on ecological sustainability. Future work can explore specific models for predicting such savings.

6 Conclusion

This paper explores ways to support communities in creating scientific knowledge by providing two systems; one supports communities in working with experts while the other supports motivated communities in acquiring procedural knowhow to design and run experiments. Our results indicate that the success of these systems depends on the motivation and capacity of the communities and supporting organizations.

ACKNOWLEDGMENTS

This work was supported in part by the Ataxia-Telangiectasia Children's Project and by NIH grant 1R01CA204585-01 as part of the NSF/NIH Smart and Connected Health program. We thank all participants who used Hevelius and Gut Instinct. We thank Nergis and Winnie from Massachusetts General Hospital, and Sarah and Jen from A-T Children's Project for their invaluable help in communicating with the A-T community members. Additionally, Vineet thanks Chen, Dingmei, Liby, Kaung, Orr, and Tushar for their help developing the Gut Instinct system; members of UC San Diego Design Lab for their feedback; and Adriana and Austin for running the kombucha experiment.

REFERENCES

[1] National Academies of Sciences , Engineering, and Medicine. 2018. *Learning through citizen science: enhancing opportunities by design*. National Academies Press.

[2] Tim Althoff, Rok Sosič, Jennifer L. Hicks, Abby C. King, Scott L. Delp, and Jure Leskovec. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547, 7663: 336–339.

[3] Audubon. 2016. Audubon Science. Using data to realize the best conservation outcomes. Retrieved December 31, 2016 from http://www.audubon.org/conservation/science/christmas -bird-count

[4] Rick Bonney, Caren B. Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V. Rosenberg, and Jennifer Shirk. 2009. Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. *BioScience* 59, 11: 977–984. https://doi.org/10.1525/bio.2009.59.11.9

[5] Project BudBurst Boulder Colorado. 2016. Project BudBurst: An online database of plant phenological observations. Retrieved December 31, 2016 from http://budburst.org/

[6] Justin Cheng and Michael Bernstein. 2014. Catalyst: triggering collective action with thresholds. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, 1211–1221.

[7] Stephanie N Chilton, Jeremy P Burton, and Gregor Reid. 2015. Inclusion of fermented foods in food guides around the world. *Nutrients* 7, 1: 390–404.

[8] Seth Cooper, Firas Khatib, Adrien Treuille, and Et Al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307: 756–760.

[9] Michael J. Coren and Fast Company. 2011. Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks. Retrieved December 31, 2016 from https://www.scientificamerican.com/article/foldit-gamers-solve-riddle/

[10] Lorenzo Coviello, Yunkyu Sohn, Adam D.I. Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A. Christakis, and James H. Fowler. 2014. Detecting emotional contagion in massive social networks. *PLoS ONE*.

[11] Dana Lewis. Real-World Use of Open Source Artificial Pancreas Systems. Retrieved from https://openaps.org/2016/06/11/real-world-use-of-open-source-artificial-pancreas-systems-poster-presented-at-american-diabetes-association-scientific-sessions/

[12] E Ernst. 2003. Kombucha: a systematic review of the clinical evidence. *Complementary Medicine Research* 10, 2: 85–87.

[13] f.lux. 2019. f.lux: sleep research. Retrieved from justgetflux.com/research.html

[14] Krzysztof Z Gajos, Katharina Reinecke, Mary Donovan, Christopher D Stephen, Albert Y Hung, Jeremy D Schmahmann, and Anoopum S Gupta. 2020. Computer mouse use captures ataxia and parkinsonism, enabling accurate measurement and detection. *Movement disorders : official journal of the Movement Disorder Society* 35, 2: 354–358. https://doi.org/10.1002/mds.27915

[15] John D Hansen and Justin Reich. 2015. Democratizing education? Examining access and usage patterns in massive open online courses. *Science* 350, 6265: 1245–1248.

[16] Eric von Hippel. 2005. Democratizing innovation: The evolving phenomenon of user innovation. *Journal fur Betriebswirtschaft* 55, 1: 63–78. https://doi.org/10.1007/s11301-004-0002-8

[17] Maia Jacobs, Galina Gheihman, Krzysztof Z Gajos, and Anoopum S Gupta. 2019. "I Think We Know More than Our Doctors": How Primary Caregivers Manage Care Teams with Limited Disease-Related Expertise. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW. https://doi.org/10.1145/3359261

[18] Alexander Kawrykow, Gary Roumanis, Alfred Kam, Daniel Kwak, Clarence Leung, Chu Wu, Eleyine Zarour, Luis Sarmenta, Mathieu Blanchette, and Jérôme Waldispühl. 2012. Phylo: A citizen science approach for improving multiple sequence alignment. *PLoS ONE* 7, 3. https://doi.org/10.1371/journal.pone.0031362

[19] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 23–34.

[20] Jeehyung Lee, Wipapat Kladwang, Minjae Lee, Daniel Cantu, Martin Azizyan, Hanjoo Kim, Alex Limpaecher, Snehal Gaikwad, Sungroh Yoon, Adrien Treuille, and Rhiju Das. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6: 2122–2127.

[21] Wendy E Mackay, Caroline Appert, Michel Beaudouin-Lafon, Olivier Chapuis, Yangzhou Du, Jean-Daniel Fekete, and Yves Guiard. 2007. Touchstone: exploratory design of experiments. *CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing System*: 1425–1434.

[22] D. W. Martin. 2007. *Doing psychology experiments*. Cengage Learning.

[23] Daniel McDonald, Embriette Hyde, Justine W Debelius, James T Morton, Antonio Gonzalez, Gail Ackermann, Alexander A Aksenov, Bahar Behsaz, Caitriona Brennan, and Yingfeng Chen. 2018. American Gut: an Open Platform for Citizen Science Microbiome Research. *mSystems* 3, 3: e00031-18.

[24] Matthew Might. Hunting down my son's killer. Retrieved from http://matt.might.net/articles/my-sons-killer/

[25] Vineet Pandey. 2019. Citizen-led Work using Social Computing and Procedural Guidance. Doctoral Dissertation, UC San Diego. Retrieved from https://escholarship.org/uc/item/2ck2p697

[26] Vineet Pandey, Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2018. Docent: transforming personal intuitions to scientific hypotheses through content learning and process training. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, 9.

[27] Birgit Penzenstadler, Jason Plojo, Marinela Sanchez, Ruben Marin, Lam Tran, and Jayden Khakurel. 2018. The DIY Resilient Smart Garden Kit. In Proceedings of the Workshop on Computing within Limits (LIMITS), Calgary, AB, Canada, 6–7.

[28] Barath Raghavan and Daniel Pargman. 2016. Refactoring Society : Systems Complexity in an Age of Limits.

[29] Katharina Reinecke, Ann Arbor, and Krzysztof Z Gajos. 2015. LabintheWild : Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

[30] Katharina Reinecke and Krzysztof Z Gajos. 2014. Quantifying visual preferences around the world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 11–20.

[31] Bethany Rittle-Johnson and Martha Wagner Alibali. 1999. Conceptual and procedural knowledge of mathematics: Does one lead to the other? *Journal of Educational Psychology* 91, 1: 175–189.

[32] Elena Rocco. 1998. Trust Breaks down in Electronic Contexts but Can Be Repaired by Some Initial Face-to-Face Contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '98), 496–502. https://doi.org/10.1145/274644.274711

[33] Kentaro Toyama. 2015. *Geek heresy : rescuing social change from the cult of technology / Kentaro Toyama*. PublicAffairs, New York.

[34] UNESCO. 2016. Nearly 69 million new teachers needed to achieve global education goals, UNESCO reports. Retrieved from https://news.un.org/en/story/2016/10/541902-nearly-69-million-new-teachers-needed-achieve-global-education-goals-unesco

[35] Paul Wicks, Timothy E Vaughan, Michael P Massagli, and James Heywood. 2011. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology* 29, 5: 411–414.

[36] Wikipedia. Community Insights/2018 Report. *2018*. Retrieved from https://meta.wikimedia.org/wiki/Community_Insights/2018_Report#What_progress_has_been_made_in_the_diversity_of_Wikimedia_communities

[37] Wikipedia. 2018. Bristol stool scale. Retrieved from en.wikipedia.org/wiki/Bristol_stool_scale

[38] World Health Organization. 2018. Dangers of poor quality health care revealed 'in all countries': WHO report. Retrieved from https://news.un.org/en/story/2018/07/1013942

[39] Zooniverse. 2007. Galaxy Zoo. Retrieved December 31, 2016 from galaxyzoo.org