Galileo: An online learning system for people to design, review, and run experiments

Vineet Pandey[1], Tushar Koul[1], Chen Yang[1], Daniel McDonald[2], Rob Knight[2], Scott Klemmer[1]
[1]Design Lab, [2]Department of Pediatrics
UC San Diego, La Jolla, CA
{vipandey, tkoul, chy099, danielmcdonald, robknight, srk}@ucsd.edu

ABSTRACT
People have intuitions and folk theories about their health and lifestyle but lack the expertise to systematically test them. This paper introduces a citizen-driven approach for people to investigate their questions. This paper investigates the power of procedural learning to help online learners with complex, creative tasks such as generating novel, structurally-sound experiments. This approach is instantiated in the *Galileo* social computing system. Using Galileo, end users design scientific experiments, improve them via reviews, and run them with other participants. Galileo's procedural learning employs templates, examples, and checklists. This paper reports on two empirical investigations of Galileo's approach. A between-subjects experiment with 72 participants found that Galileo's scaffolding significantly improved novices' experimental designs. Remarkably, participants created better experiments than doctoral students trained in experimental design. 17 volunteer health enthusiasts tested Galileo in the wild by participating in a week-long experiment. Compliance to experimental steps and correctness of data emerged as important future challenges.

Author Keywords
Online learning; social computing systems; citizen science; crowdsourcing

ACM Classification Keywords
K.3.1. [Computer Uses in Education]: Distance learning, Collaborative learning

THE PROMISE OF END USERS LEADING EXPERIMENTS
People around the world participate in experiments as data donors: browsing online [7], using activity trackers, or by joining scientific projects. Such large-scale collection of data has enabled scientific discoveries and generated valuable insights around activity inequality [1] and people's aesthetic preferences [31]. For example, scientists working with 23andme have published 94 papers since 2011 on questions such as how our genome makes us susceptible to certain conditions like *misophonia*, strong aversive reaction to others' loud chewing noises. Similarly, the American Gut Project has learned that plant diversity in food intake is an important predictor of one's gut microbiome [24] and the *f.lux* color temperature software is running experiments on the effect of blue light on sleep quality (justgetflux.com/research.html). In addition to the role that internet plays in furthering professional science, how might we bring more expertise to people so they can answer *personally-meaningful* questions?

While professional scientists and commercial ventures run experiments every day, the context and motivational pressures are different for them than for end users. This biases which questions get asked, what studies get run, and what knowledge gets created. End users may have different goals and values. This omission of more stakeholders' perspectives has practical consequences. For example, limited resources and economic motivations can focus scientific and commercial endeavors on stuff that affects lots of people; niche communities are often overlooked. One avenue that some overlooked communities have explored is political pressure and activism (*e.g.*, Lyme Disease [21]). Another approach is community-created toolkits; Type 1 diabetes community has done this effectively with monitoring tools [27].

Scientific Collaboration Online
Perhaps the most remarkable example of how global-scale collaboration can push the limits of scientific knowledge is Project Polymath (projectpolymath.org, 2009). Polymath engages mathematicians of all levels—from Fields medal winner Terrence Tao to university juniors—in proposing and collaboratively solving math problems. Foldit and EteRNA show how carefully-constructed interfaces provide novices with the micro-expertise needed to solve problems that only experts previously could [6,18]. While theoretical problems like these benefit from the creative diversity of many participants, they don't directly leverage people's lived experiences. This paper introduces strategies for leveraging people's lived experiences in ways that complement more traditional science.

This paper contributes a social computing system that enables novices to design experiments, review them with a community, and run them to generate evidence for/against their intuition. Galileo's primary architectural insight is to explicitly integrate procedural learning to *scaffold the doing and not just the knowing*.

This paper reports on two empirical investigations of Galileo's approach. A between-subjects experiment with 72 participants found that Galileo's scaffolding significantly improved novices' experimental designs. Remarkably, participants created better experiments than doctoral students trained in experimental design. 17 volunteer health enthusiasts tested Galileo in the wild by participating in a week-long experiment. Compliance to experimental steps and correctness of data emerged as important challenges to tackle.

Complementing Global Data Collection with Distribution of Expertise

Lead users can design better or at least, usefully different products from experts [11]. Snowboarders improve their binding ergonomics and Type 1 diabetes patients develop Continuous Glucose Monitoring tools. Lead-user innovation is most successful when two conditions are present. The first is contextual domain knowledge. Second, lead-user innovations—even more than expert ones—are aided by rapid feedback by using the designs.

The gap between gut instincts and feedback widens for domains where multiple factors can interfere (such as health and well-being) and make causality harder to detect. While almost everyone has intuitions about health and lifestyle, few have supporting evidence. Consequently, lead-user innovation has been less successful in these complex socio-technical domains. To be fair, the experts often struggle in these settings too.

Self-tracking Offers Insights but Not Causality

Personal objectives can provide a motivating setting for scientific work [12]. People suffering from ailments frequently track different inputs and symptoms [25]. However, many fail to find much value; common concerns include tracking too much, measurement when multiple things change simultaneously, prematurely abandoned efforts, the biases inherent in self-report, spurious correlations, and incorrect analyses [5,19]. Worse still, people falsely believe that when one event follows another, the initial event is the cause: *post-hoc ergo propter hoc*.

Expert Support Helps Find Insights, but Scaling is Hard

In some cases, expert mentorship has enabled passionate lead-users to demonstrate novel causal mechanisms (or findings) about people's health. Platform support by patientslikeme.com and scientists in the community enabled ALS patients to disprove a claim in a scientific paper that consuming lithium alleviated ALS symptoms [37]; these findings were shown to be false in a subsequent study by university researchers too. Extreme daily journaling plus intense patient-doctor relationship enable a patient to rid themselves of migraines by systematically trying different treatments including medicines and botox over four years. [9]. More recently, clinicians, researchers, and patients have collaborated using apps to discover insights about Irritable Bowel Syndrome [14]. Such systems invariably don't scale

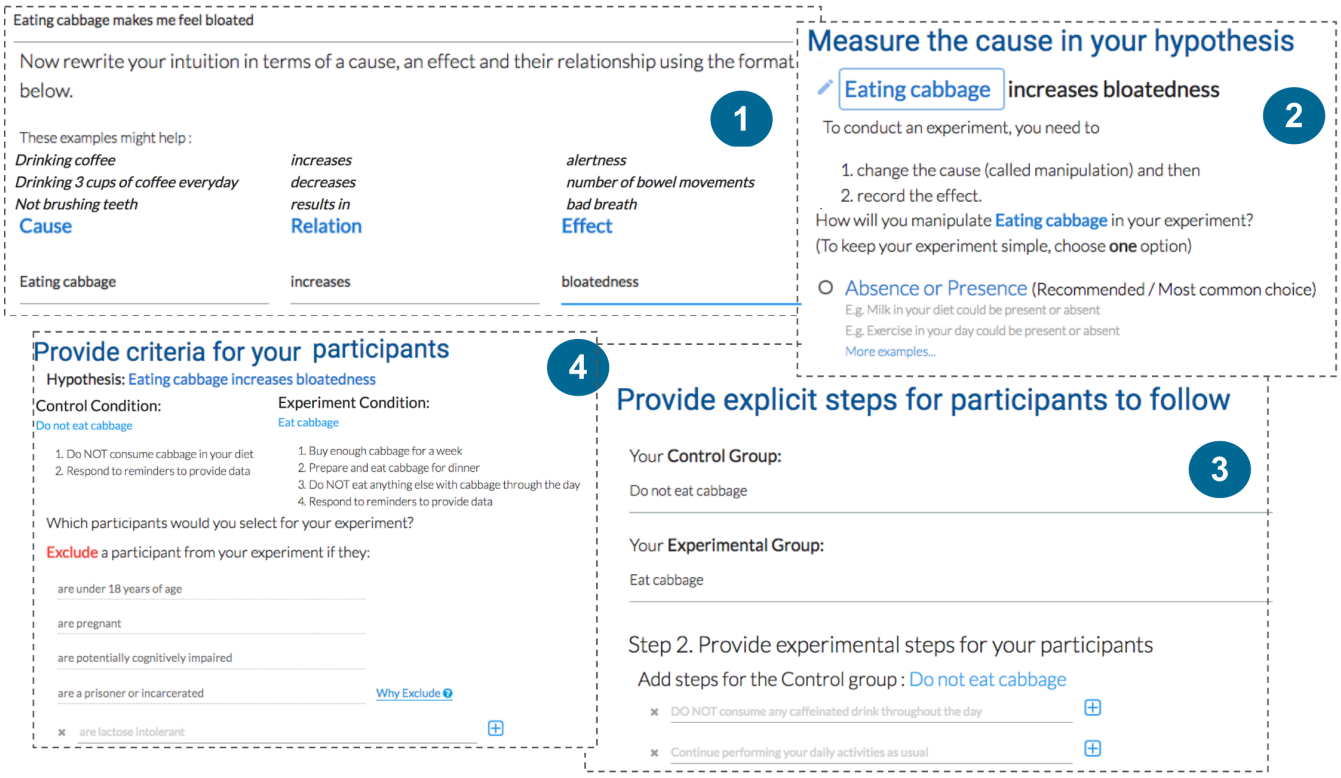

Figure 1: Design module enables people to transform their intuition to an experimental design. People (1) convert their intuition to a hypothesis, (2) provide ways to manipulate/measure cause and effect, (3) add and review control and experimental conditions and (4) provide inclusion/exclusion criteria

up since they rely on the support of a core group of scientists and clinicians who are already busy.

Aligning the Scientific Method with Personal Objectives

The scientific method—hypothesis-driven controlled experiments—can help distinguish superstitions from reusable insights. The scientific method is specifically designed to counter biases and to ensure correctness and reliability of evidence for general applicability. Even experts sometimes struggle to perform every scientific activity correctly: selecting the correct experimental design, designing it correctly, and running it successfully [23]. Research systems, like Touchstone [22], have simplified design and analysis of experiments for researchers. Some tools, like Planout [2], enable experimenters to rapidly tweak configurations in online field experiments. Such systems support people who already have expertise in experimental design. How can we train people in designing and running experiments to answer their personally-meaningful questions?

THE GALILEO EXPERIMENTATION SYSTEM

Galileo introduces a platform for motivated end users to design experiments, get them reviewed by a community, and run them with interested participants. Galileo provides process training at different steps, an online collaboration platform, and email and text-based reminders. Galileo was designed via multiple iterations with early and lead users. Early participants provided in-person feedback about the ease of using the interface. Later participants designed and ran experiments to provide feedback along with usage data that led to a number of improvements. For instance, some pilot participants did not report measurements because of the friction of logging in; consequently, Galileo now allows reports using text messages.

The Galileo web application uses the Meteor (meteor.com) framework for synchronization, Jade for the front end (jade-lang.com), Materialize for styling (materializecss.com), and Twilio as the text message gateway (twilio.com). Galileo is open for use at <URL>; its open source is at <URL>.

Design-Review-Run: From Intuitions to Investigations

This section describes Galileo's Design-Review-Run workflow.

Design: Create an experimental design from an intuition

Designing an experiment is a creative open-ended task without one correct answer [23]. As people often have many hypotheses—most of which are usually poorly-framed—providing feedback on experimental designs in the absence of experts is near-impossible. Galileo works around this challenge by providing templates and examples that help people structure experimental designs. This structured procedural training can improve on-task performance [13].

Because Galileo users are unlikely to have prior experimental design experience, Galileo walks users through a clear structured workflow. People design an experiment in Galileo by: a) converting a vague intuition to a specific hypothesis; b) providing ways to manipulate cause and measure the effect; c) providing experimental steps for control and experimental conditions; and d) providing inclusion and exclusion criteria for participants (*Figure 1*).

Many people are drawn to the informal learning of discussing their ideas online [16] and are often less keen about more formal learning [13]. While Galileo provides factual learning material, it focuses on embedding procedural learning in the interface. Once the designer has created the experiment, a final review step is performed before continuing; this provides an overview.

Review: Improve the Design via Feedback from Others

Experimental designs need to be reviewed by at least two people before they can be run. These reviewers might be friends, peers, or anyone else who can provide useful feedback. Reviewers provide both binary rubric assessments and written feedback on specific questions for each (*Figure 2*). This review module fosters iterative improvement by

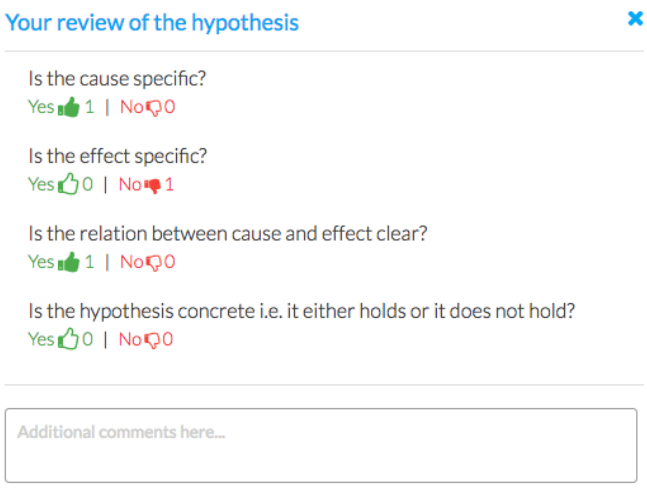

Figure 2: Participants answer specific questions about different sections of the experiment by answering questions in a rubric before providing more open-ended comments.

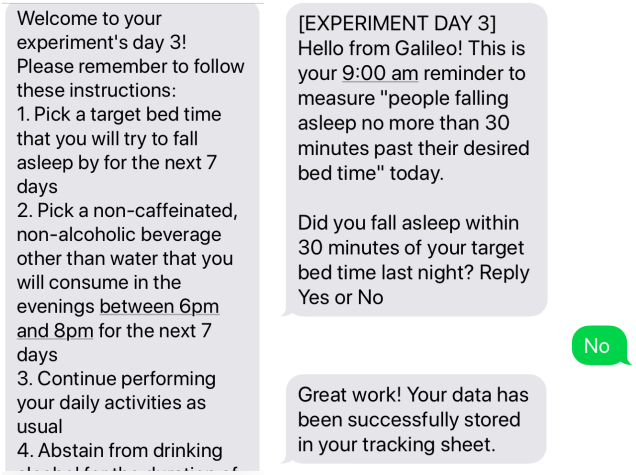

Figure 3: Galileo sends (A) A daily reminder for the experimental tasks, and (B) a request for response on compliance and measures.

enabling others to check for common errors that the designer might have missed, *e.g.*, accounting for confounds or tracking data correctly.

Run: Recruit Participants and Run your Experiment

To commence an experiment, the designer shares its unique URL with potential participants. These respondents answer questions that test whether they meet the inclusion criteria and avoid the exclusion criteria. Participants are told about the objective of the experiment but not the specific hypothesis or any experimental details apart from the steps that they need to follow. Participants explicitly consent to joining the experiment. Once underway, Galileo sends condition-specific text messages to all participants: a beginning and end of experiment message, a daily reminder, and daily cause and effect messages (*Figure 3*). People provide experimental data by responding to the cause and effect messages, which are automatically logged to the database and shown to participants in the online interface. At the end of expeirments, people received an email with a summary of the results.

FROM *KNOWING* TO *DOING*

To help novice participants design sound experiments, Galileo scaffolds *procedural* doing rather than just *conceptual* knowing: it decomposes the design into chunks, and offloads tasks like randomized placement of participants and sending data entry messages to the system. Three main insights inform Galileo's design, review, and run modules.

Manage Interdependencies to Make a Plan

Tasks like experimental design are complex in part because the work done in each step depends on choices made in a prior step. For example, the exclusion criteria for an experiment depend on the steps that participants must perform. It is likely unwise for patients with severe allergies (*e.g.*, to peanuts) to enroll in an experiment manipulating that allergen. Furthermore, the experimental steps themselves depend on the cause being manipulated. Galileo mitigates interdependency problems in two ways. First, Galileo attempts to place highly interdependent steps near each other. Second, Galileo shows the interdependencies only when necessary to enforce experimental completeness and consistency (*Figure 4*).

System Principle 1: Adjacency Mitigates Interdependencies

Working memory is heavily taxed when learning a new topic/skill. Consequently, having to recall work done in previous steps or interruptions due to moving back and forth can be especially taxing [35]. To minimize working memory, the Galileo interface groups related items. For example, to maintain the internal validity of the experiment, people need to collect data as per their experiment guidelines. Galileo asks the designer to provide details for the data collection reminders to be sent to participants right after the user decides on the cause and effect metrics (*Figure 4A*). Traditionally, these details are requested after the experimental design is complete and the logistics of the experiment are being decided.

System Principle 2: The Right Information at the Right Step

Dependencies among experimental steps cannot be entirely eliminated. Galileo manages interdependencies by presenting the right information at the right step so people can focus on the immediate step without having to explicitly manage the links between different steps. *Bridges* improve internal consistency by displaying the choices made in a previous step when necessary (*Figure 4B*). *Constraints* limit options at a step based on choices made previously so that work done at a step in an experimental design does not contradict work done at a previous step. Galileo maintains a dependency graph of different sections of an experimental

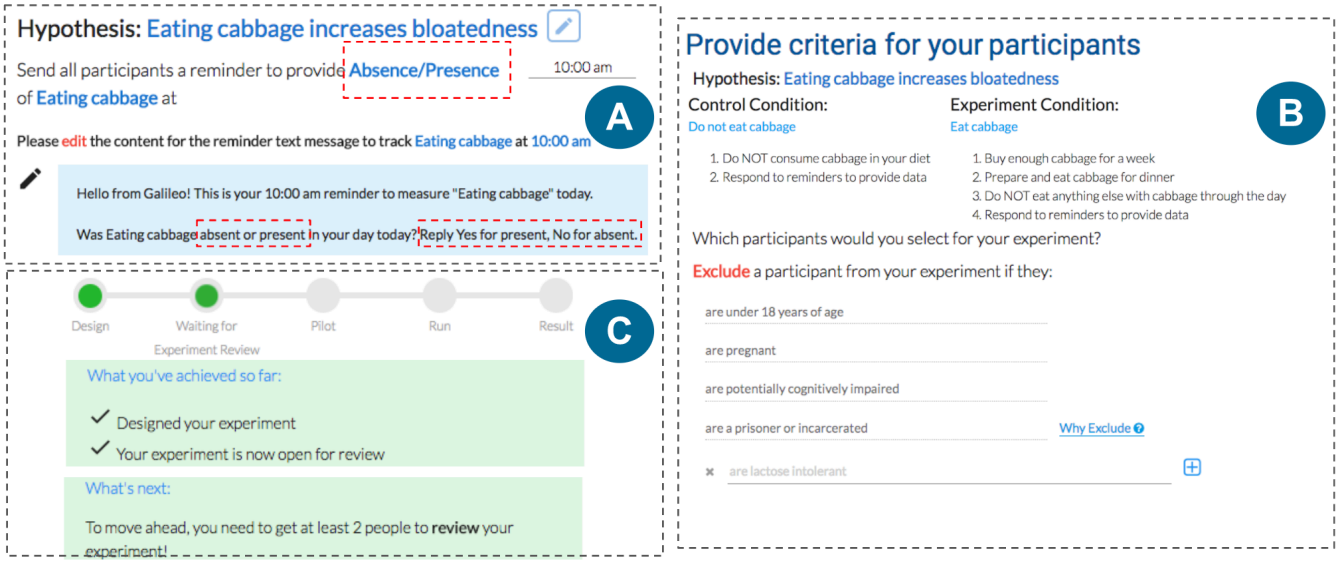

Figure 4: Managing interdependencies in experimental design. (A) Reduce interdependencies: Designers configure automatically generated data notification reminders immediately after they select how to manipulate the cause. People can edit the text messages (B) A Bridge: When people add inclusion/exclusion criteria for experiment participants, they are explicitly shown the experimental steps (C) Galileo provides specific instructions at each state to avoid overloading the user with tasks.

design to use the bridges and constraints at every step. Bridges and constraints are nudges: users can still make their own choices by ignoring the information raised by bridges or superseding the limits imposed by constraints.

System Principle 3: Automated Support Helps Run Experiments

Running an experiment requires enforcing standard procedures for all participants. Galileo performs many such tasks automatically to reduce both the designer's workload and to avoid potential biases. This consistency mirrors how researchers follow a study protocol to perform the same tasks with different participants. When people sign up to participate in a Galileo experiment, they are provided with a list of criteria set by the experiment designer; their responses automatically qualify or disqualify them for that experiment. Participants should also stay oblivious to the experiment's purpose and its internal details since it might bias them. Galileo automatically enforces this constraint by providing different views to different groups: reviewers see all the details while participants are only provided information about the steps that they need to follow. Finally, Galileo automatically places participants randomly in control and experimental groups when they join an experiment. Random assignment helps experimenters draw causal conclusions [23]. To deter overzealous designers or reviewers, Galileo enforces constraints such as not allowing the designer to review his/her experiment or for people to review and participate in the same experiment. Finally, participants receive automated texts from Galileo for the duration of the experiment with no work required from the designer.

Optimize Local Steps to Implement the Plan

Books and online learning material can overwhelm all but the most motivated [29]. Therefore, Galileo provides a learning-*by*-doing approach as opposed to learning-*then*-doing. Galileo enforces that every designer start with an intuition and then helps them convert it to a hypothesis and an experimental design. Errors in designing an experiment can come from many sources, such as having a vague hypothesis; adding confounds; or having poor data collection methods [4]. Galileo demystifies different parts of experimental design by providing templates and examples, reducing jargon, and enabling people to continuously iterate on their work (*Figure 5*).

System Principle 4: Dejargonizatorifaction

Scientists use terms well understood in their community to save time and space; however, this jargon presents a barrier for novices. In our early pilot studies, we noticed that scientific terms confused people and made them question their competence. Galileo therefore provides a simple description of a concept in common English and shows examples. For example, rather than using terms like *internal validity*, Galileo asks people to ensure that their experiment indeed tests their hypothesis by ensuring that instructions in control and experimental groups differ in exactly one step where the cause is manipulated, and by enabling people to collect appropriate data. In other cases, the scientific terms are replaced with more accessible terms. For instance, the terms *independent* and *dependent* variables even confuse some undergraduates who have taken an experimental design class. Galileo instead uses *cause* and *effect*, and provides multiple examples (*Figure 5A*). Galileo's choice of using less jargon is consistent with prior lessons that when people are focused on meeting their goals, they care less about scientific details [13]. Removing scientific jargon can potentially make Galileo more useful for a diverse online population.

System Principle 5: Bias Interface towards Doing

Writing instructions that will be followed consistently by a

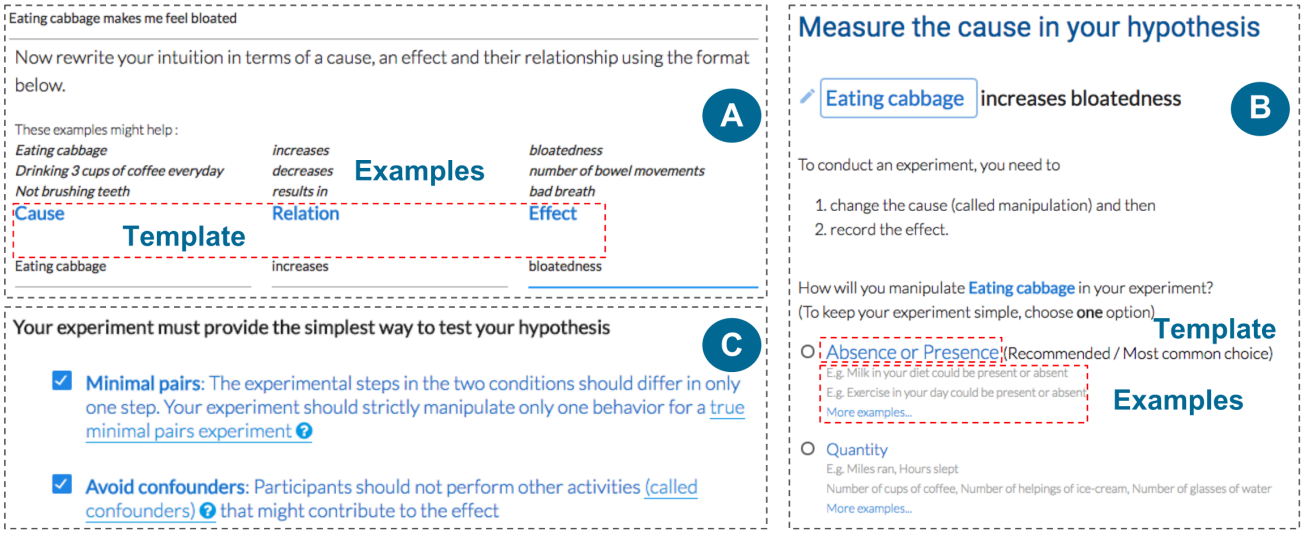

Figure 5: Optimizing local steps. (A) Reduce jargons: Galileo uses the terms cause and effect in place of independent & dependent variables. (B) Use using templates and examples to get started: Participants manipulate the cause in their hypotheses using one of the options provided. Examples provide a clear reference to how these options can be used. (C) Support iterations: Checklists at the end of every step press designers to improve their work.

diverse audience is hard. We saw this in Galileo's pilots, where people created vague experimental steps. Galileo employs templates, examples, and good defaults to create an interface that nudges people into performing specific work at every step. *Figure 5A,B* provide examples of such templates and examples. For extra-motivated learners, Galileo provides hints to improve conceptual understanding and a video overview of the concepts at the bottom; concepts are also explained in help boxes. Examples need to satisfy some criteria: they should be diverse and specific to the experimental context. Galileo enables community leaders to choose examples.

System Principle 6: Support Iterations over Pre-task Training

While examples help people tackle the cold start problem, performing work that is correct for their specific hypothesis is still a challenge. Thoughtless copying of examples does not help. To this end, instead of telling people all the requirements at the start of a step, Galileo enables people to perform some work and then iterate on this work using the checklists at the bottom. For instance, for the experimental/conditions design page, people first select the conditions, add steps and then improve steps based on the checklists at the bottom. These checklists refer to more context-specific challenges of making the experiment simple, safe, and comfortable for participants (see *Figure 5C*). This has two advantages. First, participants might be motivated by making progress rather than being bogged down by knowing beforehand all the criteria their work needs to meet. Second, multiple iterations enable people to focus on a few criteria at a time, thereby improve the quality of work. Such self-policing also provides natural learning opportunities [8]. Galileo also supports iteration at a global level. For example, many participants begin with a vague effect variable but refine it in subsequent steps as they realize they need to measure it.

EXPERIMENT: SCAFFOLDS FOR BETTER EXPERIMENTAL DESIGN

We tested the following hypotheses for designing personally-relevant scientific experiments.

Hypotheses

H1. Procedural training improves experimental design quality.

Learning research, especially online, has focused on evaluating learners' understanding of facts. The appeal of procedural learning that teaches approaches is that it helps people solve novel problems of a similar style to known ones but with different facts. The canonical domain for studying procedural learning in the literature is K-12 mathematics instruction [32]. This paper uses procedural learning with the aim of helping people generate novel, structurally-sound experiments. Effective experiments require compliance with several principles, such as a minimal-pairs design and that the measures match the manipulation. We test whether online videos are sufficient (and that procedural learning

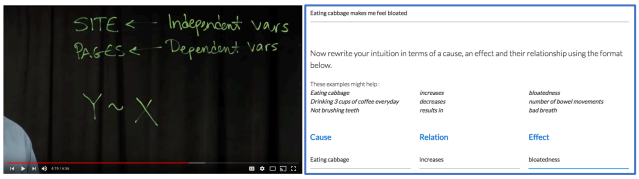

Figure 6: Two conditions for experiment. (A) *Video* condition where participants accessed videos about experimental design from a Coursera class [38] **(B) *Galileo* condition where participants accessed Galileo tool**

offers no significant benefit), and whether participants can in fact make use of the procedural training when their conceptual knowledge is still nascent [32]. An experimental design is deemed to be of good quality if it provides all relevant details to test the hypothesis (Table 1).

H2. Novices with procedural training create experiments of comparable quality to experts without procedural training.

The second hypothesis evaluates the extent to which experts' often tacit knowledge can be drawn out as explicit knowledge that improves novices' performance.

A between-subjects experiment compared the experimental design and conceptual understanding performance of participants across two conditions: *Videos* and *Galileo* (*Figure 6*). In the Videos condition, participants were provided access to online videos about experimental design, curated from a Coursera MOOC [38]. These videos were chosen since they are more interactive than others, operationalize specific concepts rather than talking vaguely about the scientific method. These videos are also among the most popular Coursera lectures. In the Galileo condition, participants were provided access to Galileo.

Method

Participants were randomly assigned to one of the two conditions. Each condition comprised an individual lab session, during which participants were asked to design an experiment for their personal intuition. An experimental design was suggested to have the following elements: a hypothesis and a possible explanation for it, ways to manipulate and measure the variables, explicit steps for participants to follow, filtering criteria for experiment participants, and tracking information needed for the variables.

A researcher introduced the condition-appropriate material. Participants were told that there was no lower or upper limit on time taken during the lab study. Each session comprised the following steps: (a) accessing the consent form, (b) accessing the condition-appropriate material and designing an experiment, (c) answering a post-test (d) filling up the survey and (e) participating in a short interview. Participants could access any online material for their task apart from the condition-specific material. Since the test was for recall, participants were not allowed to use any online resources or the tool itself. The interview asked participants about confidence in their experiment design abilities and

their experience using the system. The interview was tailored to participants' behavior and their survey responses: for example, if a participant did not watch any online videos, then the interviewer enquired why. An independent rater (a professor who teaches experimental design) blind to the conditions rated each participant's experiment using the rubric in Table 1.

Participants

Recruitment: 72 participants were recruited from a Southern California University (Table 2). Participants were novices in terms of their knowledge of experimental design. Random assignment balanced experiment design expertise.

Measures

Dependent variables comprised quality of experimental design (described in Table 1); score on a summative test for recall; and time taken to design the experiment. Qualitative measures included how participants used the tool, where they faced challenges, and a post-experiment survey.

Results

Non-parametric Mann-Whitney tested the effect of access to Galileo tool (independent variable) on the dependent variables.

Quality of experimental design: Did access to Galileo impact the quality of experimental designs? The Galileo participants generated experimental designs of better quality (M=11.3) compared to Video participants (M=5.6); *Mann–Whitney* U=108, $n1$=$n2$=36, p<0.005 (*Figure 7A*). Of the top 50% of experimental designs (36), 29 were from Galileo condition. Galileo participants performed better on five out of six sections (all except hypothesis) of experimental design, by a factor of 2. *E.g.,* the average score for designing the conditions (3pts) was 1.2 for Videos condition and 2.5 for Galileo condition.

Criteria	Operationalized as
Hypothesis	**Is the hypothesis concrete? 3 points**
	Is the cause specific? **1**
	Is the relation clear? **1**
	Is the effect specific? **1**
Measurement	**Are the cause and effect properly manipulated/measured? 2 points**
	Is the cause manipulated correctly? **1**
	Is the effect measured correctly? **1**
Conditions	**Are the conditions designed correctly? 3 points**
	Is the control condition appropriate? **1**
	Is the experimental condition appropriate? **1**
	Do the conditions differ in manipulating the cause? **1**
Steps	**Are experimental steps clear? 2 points**
	For Control condition? **1**
	For Experimental condition? **1**
Criteria	**Are the participation criteria appropriate? 2 points**
	Are the exclusion criteria correct and complete? **1**
	Are the inclusion criteria correct? **1**
Run	**Can the overall experiment be run as is? 1 point**

Table 1: The experimental design quality criteria (rated as 0: no, 1: yes). The 13-point sum represents overall quality.

Conceptual understanding: Did Galileo improve participants' post-test scores? The Galileo participants (M=3.9) performed similarly to Video participants (M=3.6); *Mann–Whitney* U=610, $n1$=$n2$=36, p=0.67 (*Figure 7B*).

Time taken: The amount of time participants spent designing experiments did not significantly differ between conditions, Mann–Whitney U=734, $n1$=$n2$=36, p=0.33 two-tailed (*Figure 7C*). This implies that higher quality of experimental design wasn't a result of spending more time on task.

Post-hoc analysis showed no correlation between time spent on task and the experimental design score for both the conditions (r=0.17 for Videos condition and r=0.22 for Galileo condition). 39 of 67 participants reported prior use of online courses, varying from occasional use of online learning material to taking as much as five classes. Preliminary analysis found no effects for people's experimental design expertise, so these were excluded from further analyses.

Comparison to experts' design

How similar were the best experimental designs to those that experts would create? To assess this, four behavioral-science doctoral students designed experiments for five intuitions selected from the highest scoring experimental designs in the Galileo condition. These students had prior training in designing and running experiments and were not provided access to Galileo. They designed 15 experiments that were rated by the same rater. Did access to Galileo enable novices to create experimental designs that were similar to experts' designs? The Galileo participants generated experimental designs of better quality (M=11.3) compared to experts (M=8.9); *Mann–Whitney* U=104, $n1$=15, $n2$=36, p<0.0005 (*Figure 7D*).

Discussion

Study involved reflection and critical thinking: Participants

Nationality	USA = 37 No Answer = 6	China = 11 Others = 18
Gender	Female = 47	Male = 24
Native English	Yes = 38	No = 34
Age	18-20 = 40 21-25 = 31	26-30 = 01
Ethnicity	Asian/Pacific = 36 White = 11	Hispanic/Latino = 14 Others = 11
Undergraduate Major	Biology = 12 Cognitive Sci = 12	Psychology = 20 Others = 20
Use of online learning material	Never = 28 2-5 classes = 12	Occasional = 16 One class = 11

Table 2: Demography info for 72 participants. Some participants did not complete portions of the survey.

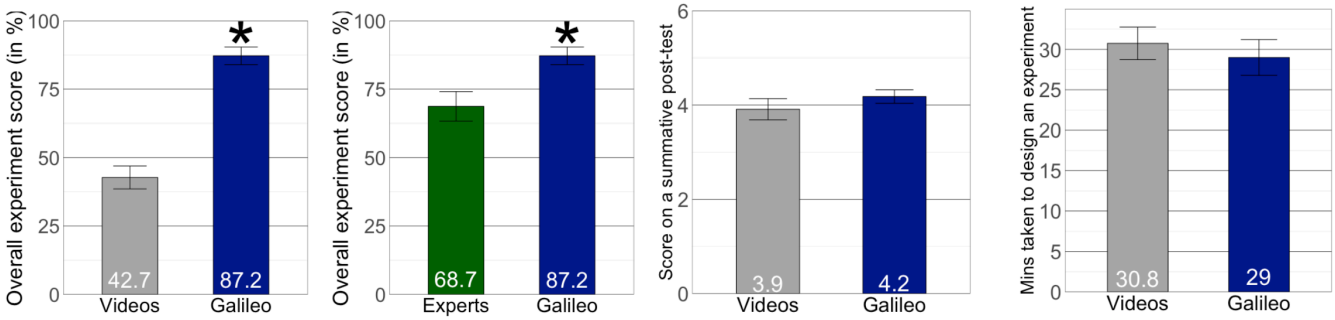

Figure 7A. *Access to Galileo improved the overall quality of experiment design

B. *Novices with Galileo created better experimental designs than experts without Galileo

C. Access to Galileo did not improve the score on a summative test

D. Access to Galileo did not alter the time spent on the experimental design task

across both the conditions mentioned that they enjoyed reflecting on their lifestyle-health ideas and thinking through how to transform their intuition into an experiment. Participants in the Galileo condition further wished that this tool was integrated with their class, describing it as "hands on" and "DIY".

Strategies used by participants about using online lectures: Participants with access to Galileo felt that the videos were slow and the interface provided them plenty of examples. Researchers found that Galileo participants opened and closed the videos in quick succession. Participants in the Videos condition, however, felt that the videos provided a refresher of some concepts they vaguely knew about. Typically, these participants followed two strategies: (1) watch all the videos at once and then begin writing the experiment, or (2) Begin designing the experiment and use the videos to fill in the gap when stuck. Since participants in Video condition created experiments of lesser quality, perhaps these strategies were less useful than Galileo's scaffolded procedural learning. Some participants in the Video condition explicitly suggested researchers add more examples to the videos.

Concern about measures: Participants in both the conditions seemed concern about their choice of measures for cause and effect. Some participants spent over 15 mins googling for the right measures to use *e.g.,* one participant looked for a formal sleep quality scale and used one from Stanford researchers.

DEPLOYMENT: HEALTH ENTHUSIASTS DESIGN AND RUN AN EXPERIMENT

18 volunteer health enthusiasts across three time zones ran a week-long experiment using Galileo. They chose to investigate whether consuming alcohol affects the time taken to fall asleep. Folk theories about the effect of alcohol on sleep abound: some participants believed that a pint of beer in the evening helps them sleep by relaxing them while others thought alcohol disturbs their sleep by interefering with their circadian rhythm; others are different ideas still. These intutions are similar to the beliefs in general American public about the effect of alcohol on sleep [28].

Design and Review: The experiment designer started with the question "Does drinking a beer in the evening help you get to bed on time?" and invited two friends to review the experiment. Both reviewers are first year graduate students who have published research designing and running experiments. The reviewers corralled the designer into creating a far more specific hypothesis: "Drinking a 5% ABV (+-0.5%) beer between 6pm and 8pm local time helps people fall asleep no more than 30 minutes past their desired bed time." (*Figure 8*)

Run: Finally, the creator reached out to friends and acquaintances who had expressed an interest in. Galileo requested all participants to adhere to the experimental protocol as much as possible without harming their health. For any confusion about the steps, participants were asked to raise a concern with the experiment designer.

28 participants expressed an interest in participating. Ten did not join: 4 failed to meet the criteria and 6 participants tried joining after the experiment start date. Of the 18 participants, 17 were from USA while 1 joined from India; all spread across three time-zones. 13 participants signed up to receive reminders by text, 5 chose email. Galileo automatically adjusts people's cause and effect receiving time based on their location. One participant stopped participating at Day 2 due to upcoming travel leading to 17 final participants (8 in control, 9 in experimental condition).

Results: This citizen experiment did not find evidence for their hypothesis. There wasn't a significant difference between the control (M=0.60) and experimental groups (M=0.64); $t(105)$=-0.34, p=0.73.

Discussion: Nudging People towards Compliance and Correctness

The overall experiment—from its design to analysis—took 11 days.

Design and Review: The experiment designer knew some concepts about experimental design but hadn't previously designed an experiment. They reported simply following

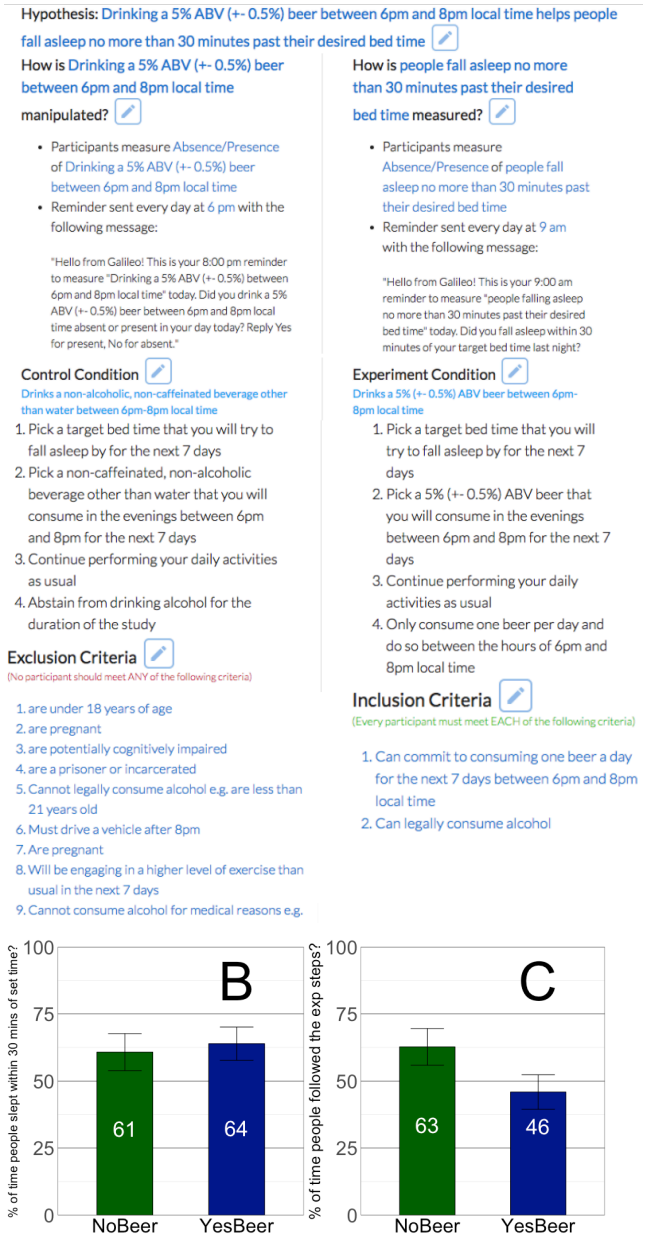

Figure 8: (A) The experiment created by a Galileo user: 17 people participated. (B) The experiment did not find evidence for this hypothesis; participants in both conditions slept within 30 mins of set time at a similar frequency. (C) Compliance emerged as an important challenge. Here, compliance was higher in the control condition.

the design workflow to create their experiment and then seeking reviews that contained specific and actionable comments. The reviewers provided helpful comments to avoid confounds by explicitly pointing out the interaction effects of caffeine with alcohol towards being unable to fall asleep on time. For some comments, the experimenter needed to operationalize reviewers' comments. Galileo can help people to provide actionable—rather than vague—feedback [26]. Adding others as collaborators with edit access can help novices receive direct edits to their work (with explanatory comments).

Compliance. Absence of evidence is different than evidence of absence: participants need to follow the experimental steps for the experiment to generate causal evidence. Did participants drink (or not drink) beer as per the experimental instructions? The compliance was low with people following the drinking instructions only 54% of the time on average (64% for control, 45% for experimental, $t(107) = 1.88, p = 0.06$). 15 of 17 participants failed to comply on at least one day**.** Moreover, some participants made tweaks to experimental steps *e.g.,* some participants preferred to drink wine rather than beer, some drank late into the night, or drank more than one beer.

This is not surprising, compliance is a real issue in experiments run by professional scientists too [23]. Additionally, Galileo does not remunerate participants, nor does it possess the authority effect of professional scientists, implying that people might be more casual about their participation. At the time of joining the experiment, participants inquired about the possibility of being placed in specific conditions hinting that they were unsure of their compliance. Galileo aims to employ two strategies to mitigate this in the future: (1) teach people of the importance of complying with the experimental conditions for generating valid results, and (2) enable the experiment designer to nudge non-compliant participants by providing a dashboard showing daily compliance and a one-click text message reminder interface.

*Latency of response***:** The quality of people's behavioral reports significantly decreases with response latency, in part because people's recall is worse and more selective than they think [36]. All participants (who responded to a post-experiment survey) favorably reviewed providing data using text messages. However, Galileo participants were busy and frequently forgot to respond to messages despite receiving and seeing the reminder text. 29% of all responses came more than 24 hours late, after additional reminders to people to provide data. Some participants traveled or went camping and mentioned feeling less inclined to check their phone or reply to messages. In such cases, citizen science tools should provide secondary ways for people to continue tracking without digital tools. Finally, background tracking using mobile or other devices (*e.g.,* number of steps walked) can ease data collection for certain experiments. Multiple participants mentioned that alongside providing their response to cause and effect reminders, they also wanted to provide more details about their activities or why they failed to comply; Galileo plans to support this in the future.

Most participants replied with significant delays in the first two days but replied faster (within 30mins or so) after the third day, hinting that forming new habits takes time. One approach to fix the cold start problem might be to remove the data from the first two days; another is to prepend a practice phase.

A dashboard for the experiment designer and the participants: Multiple participants mentioned that they wanted to view their data at a consolidated location. While citizen science systems must be careful about *when* data is revealed to participants--sometimes it may be best to wait until the experiment is complete—we agree that having good summary views and dashboards is valuable.

The experiment designer needs to be provided updates on the experiment's progress. For instance, the interface should support inputs from participants about improving the clarity of experimental steps. More specifically, the experiment designer should be provided tools to be able to send text reminders or other notifications to participants—either nudging them to provide data or providing personal messages urging them to continue.

Challenges of underpowered and or biased study: The above challenges highlight the need for tools that coach experiment designers on how best to increase compliance, accuracy, and retention. Professional scientists know how to protect against advanced challenges like social desirability bias, self-report bias, and to avoid running underpowered studies [23]. Galileo can potentially add these in the design workflow or the review or both. To glean insights from real-world settings [30], having a large set of participants is important to tackle many unsaid confounds by randomizing across different conditions.

General Discussion: Computationally-mediated Systems for Citizen Experiments at Scale

Integrating Online Communities with Experimentation in Social Computing Systems

Large scale experimentation have fruitfully improved websites [15], learning [20], and fashion [17]. Such systems support people who already have expertise in experimental design. This paper attempts to address this gap by providing a system for a community to design and run experiments for collaborative knowledge-making. Online communities, like patient health forums (*e.g.*, patientslikeme.com, lymedisease.org), can integrate experimentation with their rich discussions about health conditions . Members of a community can take up different roles for different experiments to help others and to develop expertise in designing, reviewing, and running experiments.

People are interested in better understanding their data from new technologies about personal genome and microbiome especially in clinical settings [33,34]. With systems that enable citizen-experimentation, people can potentially match scientists' knowledge with their lived experiences to create insights both for themselves and for the scientific community. Imagine if people were designing and participating in experiments by making small tweaks to their lifestyle? While Galileo's citizen experiment demonstrated issues of compliance and potential data correctness, it also showed that participants were engaged with the experiment. Most reported enjoying the process of collaborative experimentation.

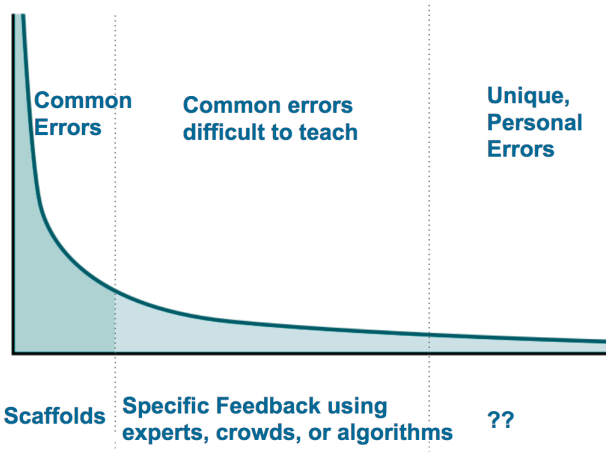

Figure 9: Scaffolds scale well and enable people to avoid making common errors. Program synthesis like techniques might generate templated feedback given error models that contain errors that are difficult to teach. Finally, correcting the long tail of unique, personal errors is much more difficult.

Opportunities for Procedural Learning at Scale

Galileo enabled participants to create useful experimental designs. We believe participants created better designs using Galileo since it provided explicit support for the different steps in experimental design process. Since the Mann-Whitney test is conservative, the possible gains may be more. Soylent introduced a 3-step crowdsourcing architecture to mitigate against common errors [3]. In a similar spirit, Galileo provides explicit guidance and support both at every step and by providing a global structure to the experimental design process.

Premade scaffolds work for frequent, anticipated challenges. Scaffolds for designing and reviewing experiments can tackle common errors by teaching people (*Figure 9*) but they don't scale for unique errors made by people; an interface or a system with infinite checklists, templates etc. won't make for a pleasing interaction. Program synthesis has shown promise in providing personalized feedback in programming assignments [10]; such techniques can be explored for experimental designs by curating error models of both popular and unique mistakes. To enable rapid iteration on the quality of experimental design, Galileo intends to explore smart routing algorithms that can invite rapid real-time feedback from a pool of community experts. Well-designed online systems bring the complementary strengths of people and algorithms; merging people's intuitions with algorithmic feedback can enable Galileo-like ideas to scale well.

CONCLUSION

This paper investigated integrating procedural learning with a complex, creative task. We demonstrate this using the *Galileo* social computing system in the context of experimental design. Using Galileo, end users designed scientific experiments, improved them via reviews, and ran them with

other participants. A between-subjects experiment with 72 participants found that Galileo's scaffolding significantly improved novices' experimental designs: participants created experiments at par with graduate students trained in experimental design. These results suggest that scaffolded procedural learning can enable novices to tackle complex tasks where conceptual knowledge might show limited benefits. Furthermore, 17 volunteer health enthusiasts tested Galileo in the wild by participating in a citizen-designed week-long experiment. This experiment illustrated the challenges of compliance to experimental steps and correctness of data. Since citizen experiments enable people to test intuitions drawn from lived experiences, they have the potential to massively scale up knowledge creation. Bringing more expertise to people can enable them to answer personally-meaningful questions that might complement traditional science.

REFERENCES

1. Tim Althoff et al. 2017. Large-scale physical activity data reveal worldwide activity inequality. *Nature* 547, 7663: 336–339. https://doi.org/10.1038/nature23018
2. Eytan Bakshy et al. 2014. Designing and deploying online field experiments. *Proceedings of the 23rd international conference on World wide web - WWW '14*: 283–292. https://doi.org/10.1145/2566486.2567967
3. Michael S. Bernstein et al. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10)*, 313–322. https://doi.org/10.1145/1866029.1866078
4. Lewis Carroll. 2006. Analysis of Errors. 1–7. Retrieved from http://faculty.sites.uci.edu/chem2l/files/2011/04/RDGerroranal.pdf
5. Eun Kyoung Choe et al. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*: 1143–1152. https://doi.org/10.1145/2556288.2557372
6. Seth Cooper et al. 2010. Predicting protein structures with a multiplayer online game. *Nature* 466, 7307: 756–760.
7. Lorenzo Coviello et al. 2014. Detecting emotional contagion in massive social networks. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0090315
8. Nigel Cross. 2004. Expertise in design: An overview. *Design Studies* 25, 5: 427–441. https://doi.org/10.1016/j.destud.2004.06.002
9. Atul Gawande. 2017. The heroism of Incremental care. *New Yorker*. Retrieved from https://www.newyorker.com/magazine/2017/01/23/the-heroism-of-incremental-care
10. Andrew Head et al. 2017. Writing Reusable Code Feedback at Scale with Mixed-Initiative Program Synthesis. In *Learning@Scale 2017*.
11. Eric von Hippel. 2005. *Democratizing innovation: The evolving phenomenon of user innovation*. MIT.
12. Charlene Jennett et al. 2016. Motivations, learning and creativity in online citizen science. *Journal of Science Communication* 15, 3.
13. Vineet Pandey Justine Debelius, Embriette R Hyde, Tomasz Kosciolek, Rob Knight, Scott Klemmer. 2018. Docent: Social computing architecture that helps people create personally-relevant scientific hypotheses. *Learning@Scale 2018*: 1–12.
14. Ravi Karkar et al. 2017. TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. *ACM Conference on Human Factors in Computing Systems*.
15. Ron Kohavi and Stefan Thomke. 2017. The Surprising Power of Online Experiments. *Harvard Business Review*. Retrieved from https://hbr.org/2017/09/the-surprising-power-of-online-experiments
16. Chinmay Kulkarni et al. 2015. Talkabout: Making distance matter with small groups in massive classes. *CSCW: ACM Conference on Computer Supported Collaborative Work*.
17. R Kumar and K Vaccaro. 2017. An experimentation engine for data-driven fashion systems. *AAAI Spring Symposium - Technical Report* SS-17-01-: 389–394.
18. Jeehyung Lee et al. 2014. RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences* 111, 6: 2122–2127. https://doi.org/10.1073/pnas.1313039111
19. Ian Li et al. 2010. A stage-based model of personal informatics systems. *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10*: 557. https://doi.org/10.1145/1753326.1753409
20. Derek Lomas et al. 2013. Optimizing Challenge in an Educational Game Using Large - Scale Design Experiments. *CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, April: 89–98. https://doi.org/10.1145/2470654.2470668
21. LymeDisease.Org. Political Action Related to Lyme Disease. Retrieved from

https://www.lymedisease.org/get-involved/take-action/support-legislation/

22. Wendy E Mackay et al. 2007. Touchstone: exploratory design of experiments. *CHI '07 Proceedings of the SIGCHI Conference on Human Factors in Computing System*: 1425–1434. https://doi.org/10.1145/1240624.1240840

23. D. W. Martin. 2007. *Doing psychology experiments*. Cengage Learning.

24. Daniel McDonald et al. 2018. American Gut: an Open Platform for Citizen-Science Microbiome Research. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2018/03/07/277970.abstract

25. D. Neff, G., & Nafus. 2016. *Self-Tracking*. MIT Press.

26. Tricia J Ngoon et al. 2018. Interactive Guidance Techniques for Improving Creative Feedback. In *CHI 2018*.

27. OpenAPS.org. #WeAreNotWaiting to reduce the burden of Type 1 diabetes. Retrieved from OpenAPS.org

28. Michael J Breus Ph.D. Alcohol and Sleep: What You Need to Know. *Psychology Today*. Retrieved from https://www.psychologytoday.com/us/blog/sleep-newzzz/201801/alcohol-and-sleep-what-you-need-know

29. J. Reich. 2014. MOOC completion and retention in the context of student intent. *EDUCAUSE Review Online*.

30. Katharina Reinecke et al. 2015. LabintheWild : Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*.

31. 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.

32. 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. https://doi.org/10.1037/0022-0663.91.1.175

33. Orit Shaer et al. Informing the Design of Direct-to-Consumer Interactive Personal Genomics Reports. *J Med Internet Res* 17, 6: e146. https://doi.org/10.2196/jmir.4415

34. Orit Shaer and Oded Nov. 2014. HCI for Personal Genomics. *Interactions of ACM*. https://doi.org/10.1145/2656622

35. Cheri Speier et al. 2003. The Effects of Interruptions, Task Complexity, and Information Presentation on Computer-Supported Decision-Making Performance. *Decision Sciences* 34, 4: 771–797. https://doi.org/10.1111/j.1540-5414.2003.02292.x

36. Ronald Taft. 1954. Selective recall and memory distortion of favorable and unfavorable material. *The Journal of Abnormal and Social Psychology 49*, 23–28. https://doi.org/10.1037/h0056436

37. Paul Wicks et al. 2011. Accelerated clinical discovery using self-reported patient data collected online and a patient-matching algorithm. *Nature Biotechnology* 29, 5: 411–414.

38. Jacob Wobbrock and Scott Klemmer. 2018. Designing, Running, and Analyzing Experiments. Retrieved from https://www.coursera.org/learn/designexperiments

APPENDIX

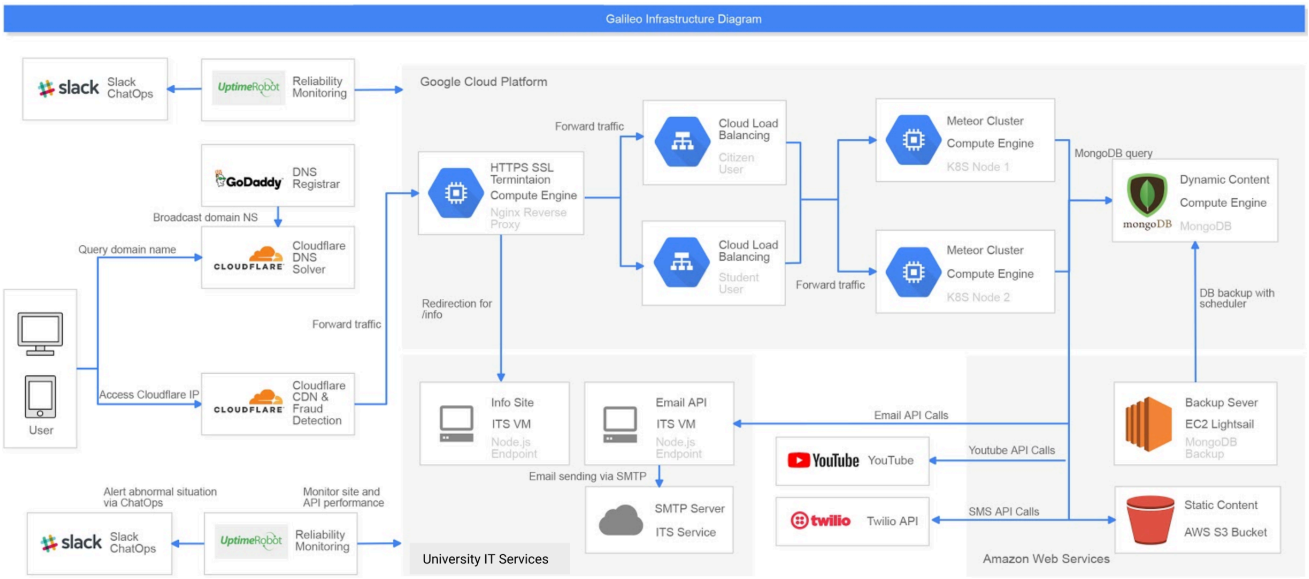

Figure 10: Galileo's cloud architecture. Galileo is hosted on Google Compute Engine with multiple integration endpoints for external services like Youtube (for hosting lectures), Twilio (for sending/receiving text messages) and University IT Services (for email calls from @university domain).