Project Sidewalk: A Web-based Crowdsourcing Tool for Collecting Sidewalk Accessibility Data at Scale

Anonymized for submission

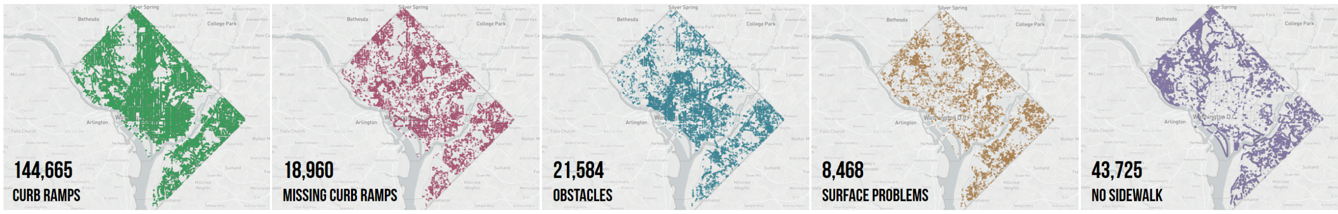

Figure 1. In an 18-month deployment study of Project Sidewalk, we collected 208,137 sidewalk accessibility labels, including curb ramps, missing curb ramps, sidewalk obstacles, and surface problems. Each dot above represents a geo-located label rendered at 50% translucency.

ABSTRACT

We introduce *Project Sidewalk*, a new web-based tool that enables online crowdworkers to remotely label pedestrian-related accessibility problems by virtually walking through city streets in Google Street View. To train, engage, and sustain users, we apply basic game design principles such as interactive onboarding, mission-based tasks, and progress dashboards. In an 18-month deployment study, 797 online users contributed 208,137 labels and audited 2,941 miles of Washington DC streets. We compare behavioral and labeling quality differences between paid crowdworkers and volunteers, investigate the effects of label type, label severity, and majority vote on accuracy, and analyze common labeling errors. To complement these findings, we report on an interview study with three key stakeholder groups (N=14) soliciting reactions to our tool and methods. Our findings demonstrate the potential of virtually auditing urban accessibility and highlight tradeoffs between scalability and quality compared to traditional approaches.

Author Keywords

Crowdsourcing; accessibility; mobility impairments; GIS

ACM Classification Keywords

H.5.m. Information interfaces and presentation *(e.g.,* HCI).

INTRODUCTION

Geographic Information Systems (GIS) such as *Google Maps*, *Waze*, and *Yelp* have transformed the way people travel and access information about the physical world. While these systems contain terabytes of data about road networks and points of interest (POIs), their information about physical accessibility is commensurately poor. GIS websites like *Axsmap.com*, *Wheelmap.org*, and *AccessTogether.org* aim to address this problem by collecting location-based accessibility information provided by volunteers (*i.e.,* crowdsourcing). While these efforts are important and commendable, their value propositions are intrinsically tied to the amount and quality of data they collect. In a recent review of accessibility-oriented GIS sites, Ding *et al.* [15] found that most suffered from serious data sparseness issues. For example, only 1.6% of the Wheelmap POIs had data entered on accessibility. One key limiting factor is the reliance on local populations with physical experience of a place for data collection. While local users who report data are likely to be reliable, the dependence on *in situ* reporting dramatically limits scalability—both *who* can supply data and *how* much data they can easily supply.

In contrast, we are exploring a different approach embodied in a new interactive tool called *Project Sidewalk* (Figure 2), which enables online crowdworkers to contribute physical-world accessibility information by *virtually* walking through city streets in Google Street View (GSV)—similar to a first-person video game. Rather than pulling solely from local populations, our potential pool of users scales to anyone with an Internet connection and a web browser. Project Sidewalk extends previous work in streetscape imagery auditing tools like *Canvas* [4], *Spotlight* [7], *BusStop CSI* [22], and *Tohme* [25], all which demonstrate the feasibility of virtual auditing and, crucially, that virtual audit data has high concordance with traditional physical audits. However, this past work has focused on small spatial regions, relied on specialized user populations such as public health researchers [4,7] and paid crowdworkers [22,25], and has not been publicly deployed.

In this paper, we present an 18-month deployment study of Project Sidewalk in Washington DC. In total, 797 users contributed 208,137 geo-located accessibility labels and virtually audited the entirety of Washington DC (1,075 miles of city streets; see Figure 1). As the first public deployment of a virtual auditing tool, our research questions are exploratory: How can we engage, train, and sustain crowd workers in virtual accessibility audits? Are there behavioral and/or labeling quality differences between paid crowd workers and volunteers? What are some common labeling mistakes and how may we correct them in future tools? Finally, how do key stakeholder groups react to our tool and what are some of their concerns?

https://doi.org/10.1145/1234567890

To address these questions, we analyzed interaction logs from our DC deployment, performed a semi-controlled data validation study, and conducted semi-structured interviews with three stakeholder groups (N=14): government officials, people with mobility impairments (MI), and caretakers. In our deployment study, we found that *registered* volunteers completed significantly more missions, on average, than our *anonymous* volunteers (M=1.5 *vs.* 5.8) and that our *paid* workers—who were compensated per mission—completed more than both (M=35.4 missions). In the data validation study, paid workers also significantly outperformed registered and anonymous volunteers in finding accessibility problems (*recall*=68% *vs.* 61% and 49%, respectively) but precision was roughly equivalent for all groups (~70%). Our findings also show that the number of issues found significantly increases with the number of labelers per street—with five labelers, recall rose from 68% to 92%.

To complement these findings, our interview study asked about perceptions of and experiences with urban accessibility and solicited reactions to Project Sidewalk and the idea of crowdsourcing accessibility in general. All three stakeholder groups were positive: while government officials emphasized cost-savings and civic engagement, the MI and caregiver groups focused more on personal utility and enhanced independence. Key concerns also arose, including data reliability, maintenance, and, for the MI participants, whether labels properly reflected their accessibility challenges (the latter echoes findings from [23]).

In summary, the contributions of this paper include: (i) Project Sidewalk, a novel web-based virtual auditing tool for collecting urban accessibility data at scale; (ii) results from an 18-month deployment and complementary data validation study exploring key behavioral and labeling quality differences between volunteer and paid crowdworkers; (iii) findings from semi-structured interviews with three stakeholder groups soliciting reactions to Project Sidewalk and identifying key concerns and design suggestions; (iv) and our open-source urban accessibility dataset. By scaling up data collection methods for sidewalk accessibility, our overarching aim is to enable the development of new accessibility-aware mapping tools (*e.g.,* [23,31]), provide increased transparency and accountability about city accessibility, and work with and complement government efforts in monitoring pedestrian infrastructure.

RELATED WORK

We present background on sidewalk accessibility, existing methods for collecting street-level accessibility data, and volunteer geographic information (VGI) systems.

Street-Level Accessibility

Accessible infrastructure has a significant impact on the independence and mobility of citizens [1,39]. In the U.S., the *Americans with Disability Act* (ADA) [54] and its revision, the *2010 ADA Standards for Accessible Design* [53], mandate that new constructions and renovations meet modern accessibility guidelines. Despite these regulations, pedestrian infrastructure remains inaccessible [17,27]. The problem is not just inaccessible public rights-of-way but a lack of reliable, comprehensive, and open information. Unlike road networks, there are no widely accepted standards governing sidewalk data (though some recent initatives are emerging [40]).

While accessible infrastructure is intended to benefit broad user populations from those with unique sensory or physical needs to people with situational impairments [57], our current focus is supporting those with ambulatory disabilities. In Project Sidewalk, we focus on five high-priority areas that impact MI pedestrians drawn from ADA standards [51–53] and prior work [34,36]: *curb ramps*, *missing curb ramps, obstacles, surface problems,* and the *lack of a sidewalk* on a pedestrian pathway.

Data Collection Methods for Street-Level Accessibility

Traditionally, collecting data on street-level accessibility has been the purview of local and state governments; however, with widespread access to the Internet and smartphones, three alternatives have emerged: *in situ* crowdsourcing where a user explicitly captures and reports data [11,15,35,37], automatic or hybrid reporting using sensors [8,28,30,43,47], and remote crowdsourcing using streetscape imagery [18,22,24,25]. Each approach has unique benefits and drawbacks—*e.g.,* in terms of data type, maintenance, and coverage—and should be considered complementary. While *in situ* crowdsourcing relies on local knowledge and is likely to produce high-quality data, both academic and commercial tools have struggled with data sparsity [15], perhaps because of high user burden and low adoption. Automatic reporting tools lower user burden by implicitly capturing accessibility data using smartphone- or wheelchair-based sensors; however, accurately converting these quantitative measurements (*e.g.,* accelerometer data) to useful sidewalk assessments is still an open research area. Moreover, these tools are limited to capturing where wheelchair users already go, not where they are *unable* to go (though [30] is attempting to address this limitation, in part, by combining sensor data with continuous video recording).

Most related to our work are virtual auditing tools of street-level accessibility using streetscape imagery. While initial research focused on establishing the reliability of GSV-based audits compared with traditional, physical-based methods [5,9,46,56], more recent work has introduced and evaluated web-based tools in controlled studies [18,22,24,25]. Project Sidewalk builds on these systems by gamifying the user experience and supporting open-world exploring via missions—similar to first-person video games. Additionally, we present the first public deployment study, which enables us to uniquely compare user behavior and labeling performance across user groups and contributes the largest open dataset on sidewalk quality in existence.

Volunteered Geographic Information (VGI)

Project Sidewalk is a new type of *volunteered geographic information* (VGI) system [21]. In VGI, non-experts

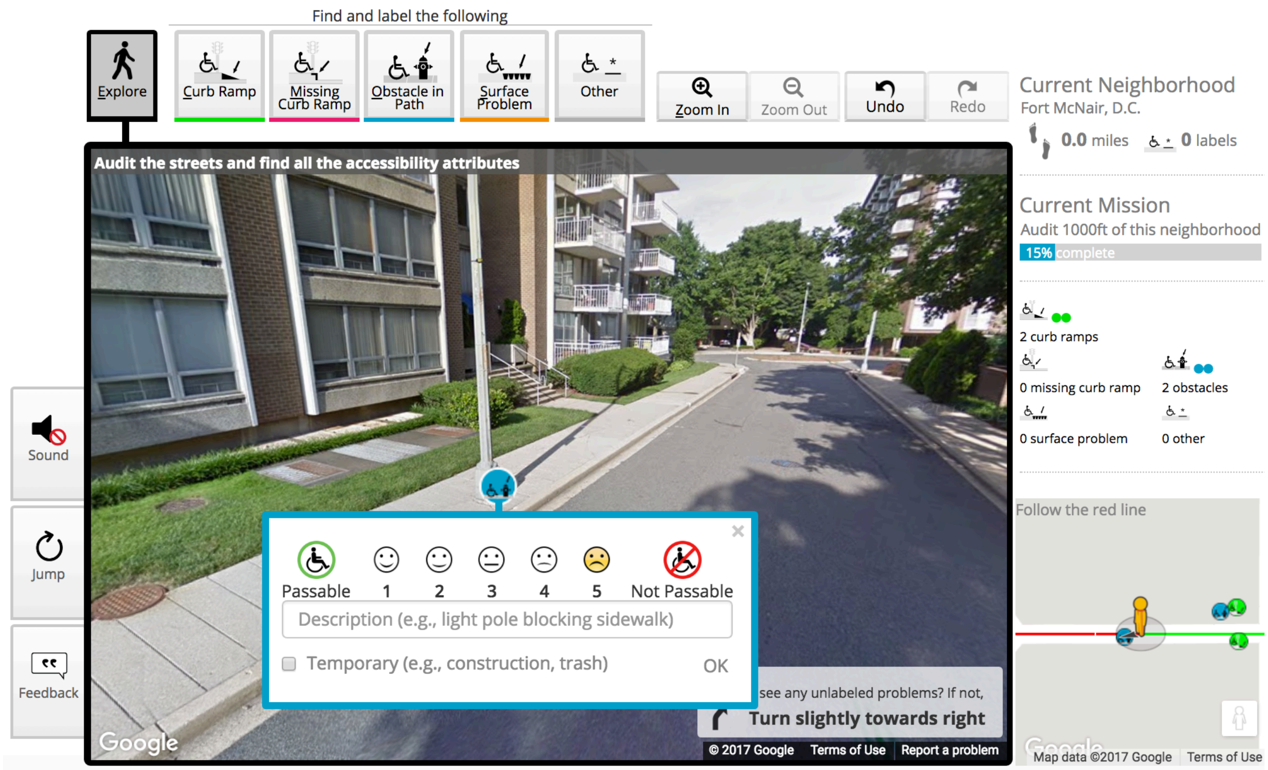

Figure 2. In Project Sidewalk, users are given missions to explore city neighborhoods and find accessibility problems. The UI is comprised of four parts: (center) GSV-based exploration and labeling pane; (top) button menu bar; (right) mission pane with progress tracking and navigation; (left) and settings menu.

contribute GIS-related data through open mapping tools—*e.g., Wikimapia*, *Mapillary*, *CycloPath* [41], and most notably, *OpenStreetMap* (OSM). In comparison to more authoritative sources, VGI data quality and spatial coverage are key concerns [3]. While some studies have shown comparable quality between VGI and government maps [19,20,33], recent work has identified strong biases in contributions correlated with population density [32,44]. We address this limitation by combining both volunteer and paid crowd workers and by eliminating the need to have physical access to a place to contribute data. Our work contributes to VGI by analyzing contribution patterns and labeling quality differences between these two user groups.

PROJECT SIDEWALK

To use Project Sidewalk, users visit projectsidewalk.io on a laptop or desktop (touchscreens are not currently supported). The landing page provides a brief description of the tool—both its purpose and how to use it—along with basic statistics and visualizations about collected data to encourage participation. Upon clicking the '*Start Mapping*' button, new users are greeted by a multi-stage interactive tutorial to learn about the user interface and basic accessibility concepts. Once the tutorial is complete, users are auto-assigned a neighborhood in DC and given their first mission. Missions guide users through specific neighborhood streets: as the user walks virtually along their route, they are asked to find, label and rate street-level accessibility issues. After completing a mission, a "mission complete" screen is displayed and a new mission is assigned. Users can choose to contribute anonymously or to register and login. We prompt anonymous users to register after finishing their first street segment. Registered users can resume missions and check their contribution activity on an interactive dashboard. Currently, however, there is no way to view or compare performance to others (*e.g.,* a leaderboard).

Training users. Training crowdworkers is difficult, especially for subjective judgment tasks like classifying entities [2]. While a wide range of training approaches are possible—from ground truth seeding with real-time performance feedback to qualification tasks that ensure proficiency [45]—our current training strategy is three-pronged. First, new users are presented with an interactive tutorial, a technique common to modern video games called *onboarding* [42]. We onboard users through an initial *guided* mission that explains the UI and key game mechanics, provides information about street-level accessibility concepts, and monitors and helps the user correct mistakes. As users step through the onboarding experience, their mission status pane is updated just like a normal mission. In total, there are 37 onboarding parts, which are designed to take less than four minutes.

Second, after completing onboarding, initial missions include pre-scripted help dialogs that are triggered based on user behavior. For example, after panning 360° around their first street intersection, Project Sidewalk helps the user use the top-down mission map to take a step in the right direction. These help dialogs are complementary to onboarding: there is an inherent tradeoff between building skills and knowledge through initial training time, and actually having the user begin using the tool in earnest.

Figure 3. Project Sidewalk has five primary color-coded label types: *curb ramps, missing curb ramps, obstacles, surface problems,* and *no sidewalk*. The images above are example accessibility issues found by users in our public deployment.

Finally, throughout every mission, our tool continuously observes user behavior and provides brief, transient usage tips to encourage proper labeling behavior and increase user efficiency. For example, if we observe that a user is not providing severity ratings, we provide a friendly reminder. If we observe only mouse clicks, we encourage keyboard shortcuts. These one-line tips auto-disappear and can also be explicitly dismissed. Importantly, we cannot provide *corrective* labeling feedback because we do not know about a label's correctness *a priori*.

Exploring and labeling. Similar to [22,25], Project Sidewalk has two modes of interaction: *explorer mode* and *labeling mode*. In explorer mode, users follow turn-by-turn directions to explore their assigned mission routes using GSV's native navigation controls. If users get lost exploring, they receive reminders to return to their mission routes, which can be clicked to auto-jump back. At any location, the user can pan (360º horizontally and 35º vertically) and zoom to assess sidewalks more closely. The user's FOV is 89.75º.

Users enter the labeling mode by clicking on a labeling button. There are five primary label types: *curb ramp, no curb ramp*, *obstacle, surface problem*, and *no sidewalk* (Figure 3). In this mode, all interactions for controlling movement and the first-person camera view (*e.g.,* pan, pitch) are disabled and the mouse cursor changes to a circular icon representing the selected label. To place a label, the user clicks directly on the accessibility target in the GSV image. A context menu then appears, which asks the user to rate problem severity on a 5-point scale where '5' represents an impassable barrier for someone in a manual wheelchair. The user can also enter additional notes in a description text field or mark a problem as temporary (*e.g.,* due to construction). After closing the context menu, Project Sidewalk automatically reverts to explorer mode.

Project Sidewalk seamlessly repositions applied labels in their correct location as the user pans or zooms—thus, labels appear to "stick" to their associated target. However, once a user takes a step, their labels are no longer visible in the GSV interface (unless they return to their original labeling location). This is due to GSV API limitations. Instead, previously placed labels can be viewed on the top-down mission map.

Missions. Missions serve a two-fold purpose: first, as a game mechanic, they provide an easy-to-understand and engaging narrative for directing data collection tasks. Second, from a system design perspective, missions provide a flexible approach to discretize, assign, and distribute work. Though we envision a variety of future mission types—*e.g.,* data validation missions, labeling user supplied imagery—our current system focuses on encouraging exploration and labeling in the GSV interface. Users are assigned a high-level goal of auditing a neighborhood and then routed on missions of increasing length and complexity within that neighborhood. Mission lengths increase from 500ft to a maximum of 0.5mi (2,640ft). Mission feedback is provided via a mission status pane, completion screens, and, for registered users, an interactive dashboard. If a user gets stuck during a mission, they can choose to "jump" to a different part of their assigned neighborhood or manually choose a new neighborhood. For finishing a mission or completing a neighborhood, users are rewarded with mission completion screens and sound effects.

IMPLEMENTATION, DATA, AND API

Creating a robust, usable, and publicly deployable system required a significant engineering and human-centered design effort. Our open-source *Github* repository has 2,747 commits from 20 team members and 43,898 lines of developed code (excluding comments). Project Sidewalk's backend is built in *Scala*, and *PostgreSQL* with the *PostGIS* spatial extension and the frontend is in *JavaScript* and *HTML/CSS*. Below, we describe four key implementation areas: preparing a city for deployment, work allocation algorithms, triangulating and clustering labels, and our API.

Preparing a city. Project Sidewalk has two data prerequisites for deployment: GSV and OSM. To construct a street network topology, we extract OSM `<way>` elements marked with street-related tags within a city's geographic boundary. We also extract `<node>` and `<nd>` elements for metadata (*e.g.,* lat-long coordinates) and links between nodes and edges. Because `<way>` polylines can extend multiple city blocks, we create smaller units, called *street segments,* by partitioning streets at each intersection. For DC, this resulted in 15,014 street segments with a total length of 1,164 miles. We filtered 892 segments that contained highways and/or where GSV imagery was unavailable due to government security precautions. In total, we were left with 14,037 segments over 1,075 miles (Figure 4).

Allocating and Distributing Work via Missions

Allocating and distributing work is a two-step process consisting of assigning neighborhoods then street segments. We use the *mission* construct to do both. We iterated on these

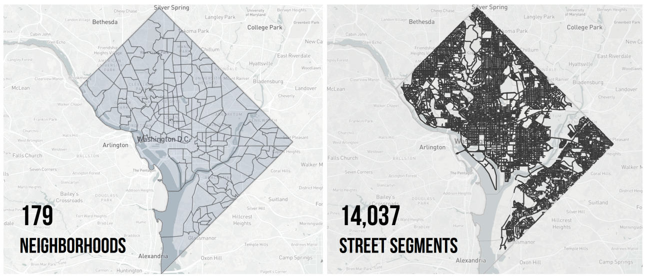

Figure 4. DC's 179 neighborhoods and 14,037 street segments (1,075mi).

task allocation algorithms throughout our deployment as we discovered inefficiencies or mistakes. Below, we present our current approach, which was used for the last three months of our deployment, and briefly mention old approaches.

Our current version is based on a "work quality" threshold determined by analyzing labeling behavior from our research group and informal manual reviews of end-user contributions. We define a "good" user as someone who contributes a minimum of 3.75 labels per 100 meters on average. While labeling frequency is an imperfect proxy for worker quality, it is easy to implement and fast to compute. We integrate this quality metric to prioritize street segments:

$$priority_{street} = \begin{cases} 1, & if\ cnt('good'\ users) = 0 \\ 1/(1+x) & otherwise \end{cases}$$

where $x = cnt(\text{"good" users}) + 0.25 * cnt(\text{"bad" users})$. This algorithm prioritizes street segments inversely proportional to the number of previous audits with a weight penalty assigned for "bad" users.

Allocating neighborhoods. Users are given missions to explore and label assigned neighborhoods. Neighborhoods are allocated at two points: after a user completes onboarding and after they complete a previously assigned neighborhood. In earlier versions of Project Sidewalk, we randomly assigned users to neighborhoods within the top ten lowest completion rates. This approach, however, treated all previous work equivalently. In the current version, we incorporate street segment priority by first calculating the mean priority of all street segments for each neighborhood and then randomly assigning neighborhoods from a list with the top five highest means. Users can also choose their own neighborhoods; however, this feature was somewhat hidden and not prominently used in our deployment.

Calculating mission routes. Mission routes are composed of street segments, which are dynamically selected when a user reaches an intersection (*i.e.,* the end of a segment). To enhance immersion and limit user confusion, the routing algorithm attempts to select contiguous segments whenever possible. In older versions of Project Sidewalk, the segment selection algorithm simply chose a randomly connected segment that the current user had not already audited. However, this failed to incorporate work completed by other users, which was inefficient. In our current implementation, for each neighborhood, we maintain a discretized list of unaudited street segment priorities (*bin size*=0.25). When a user reaches an intersection, we randomly select any unaudited connected street segment with the same discretized priority as the highest one in the neighborhood list. If none exist, we inform the user that they have completed this part of the neighborhood and automatically transport them to the highest priority remaining neighborhood street. We use a similar process for positioning users when they first begin a new neighborhood—we place them at the beginning of the highest priority street segment.

Project Sidewalk Data

In Project Sidewalk, users label streetscape panoramas projected into 3D space [16]. We need to convert these 3D-point labels to 2D lat-lng coordinates and then aggregate multiple labels for the same target into a single cluster.

3D to 2D. To obtain geo-located labels from the 3D projection, we use: (i) the panorama's 3D-point cloud data, which is obtained by LiDAR on the GSV cars; (ii) the *lng,lat* coordinate of the GSV car; and (iii) the x_{img}, y_{img} position of the label on the panorama. More specifically:

$$\begin{pmatrix} lng_{target} \\ lat_{target} \end{pmatrix} = \begin{pmatrix} lng_{GSV\_car} \\ lat_{GSV\_car} \end{pmatrix} + \begin{pmatrix} \Delta lng \\ \Delta lat \end{pmatrix}$$

where we compute $\Delta lng, \Delta lat$ by using the x_{img}, y_{img} label position on the panorama and the 3D-point cloud data to obtain the offset dx, dy, dz at x_{img}, y_{img}. The offset is in meters, which we convert to $\Delta lng, \Delta lat$ and plug into the equation.

Raw label data. For each label, we record three sets of information: who provided the label and when, how the data was collected in GSV (the user's POV, heading, source panorama id), and information about the label itself, such as *label type*, *lat-long position*, *x,y position* on panorama, *severity rating*, *textual description*, and a *temporary* flag.

Clustering. Because users can find and label the same accessibility problem from different panoramas, we needed to develop an algorithm to aggregate labels for the same target together. We do this by clustering. Each cluster refers to a single found problem (and may contain one or more raw labels). We use a two-stage clustering approach: *single-user* clustering followed by *multi-user* clustering. First, we consolidate raw labels for each individual user into intermediate clusters—this is necessary because some users choose to label a single problem from multiple viewpoints. Second, we combine these individual user clusters together to create our final cluster dataset. Both stages use the same hierarchical agglomerative clustering approach: the Vorhees clustering algorithm with the haversine formula to compute distances between labels and clusters.

For stage one, we cluster raw labels of the same type that are within a certain distance threshold. Because some label types are often legitimately close together—*e.g.,* two curb ramps on a corner—we use two different thresholds: 2 meters for curb and missing curb ramps and 7.5 meters for other label types. These thresholds were determined empirically by iteratively computing clusters at different threshold levels

	Volunteers		Turkers	Researchers	Total	Total
	Anon (*N*=384)	Registered (N=243)	(N=170)	(N=28)	Labels	Clusters*
Curb Ramp	10,631	27,144	88,554	18,336	144,665	51,098
M. Curb Ramp	1,310	3,250	13,262	1,138	18,960	7,941
Obstacle	1,105	2,827	16,154	1,498	21,854	12,993
Surf. Prob.	765	1,896	3,216	2,591	8,468	5,647
No Sidewalk	1,414	6,211	28,181	7,919	43,725	23,468
Occlusion	68	310	462	438	1,278	953
Other	92	148	1,137	34	1,411	928
Total Labels	**15,385**	**41,786**	**150,966**	**31,954**	**240,091**	**103,028**

Table 1. The total amount of data collected during our deployment. *Total clusters refers to filtered data only. All other columns are the full dataset.

from 0 to 50 meters (*step size*=1 meter) and qualitatively analyzing the results. Stage two clustering is similar but uses the centroids of stage one clusters with slightly looser thresholds (7.5 and 10 meters, respectively).

Public API

To enable the use and broader study of our collected data, we developed and released an initial public REST API (projectsidewalk.io/api). The API has three endpoint types: *labels* for obtaining raw label data, *clusters* for obtaining label clusters, and *scores*, which provide computed scores for street and neighborhood accessibility. Each API requires a lat-long bounding box to specify an area of interest for input and returns data in the *GeoJSON* format. For the score APIs, we developed a simple scoring model that incorporates the number of problem labels and returns an accessibility score between 0 and 1. Providing a robust, personalizable, and verifiable scoring algorithm is ongoing work.

DEPLOYMENT STUDY

In August of 2016, we launched an 18-month deployment study of Project Sidewalk. Washington DC was selected as the study site because of its large size (158 km^2), diverse economic and geographic characteristics, and substantial commuter population—many of whom take public transit and use pedestrian infrastructure [50]. Additionally, as the nation's capital, which draws ~20m visitors/yr [55] , there is increased pressure to follow and model ADA guidelines.

We recruited two types of users: *volunteers* through social media, blog posts, and email campaigns, and *paid crowd workers* from Amazon Mechanical Turk (turkers). We further divide volunteers into *anonymous* and *registered* groups; the former was tracked by IP address. For comparison, we also show data from 28 members of our research lab, who voluntarily contributed to help test the tool and received in-person training on *how* and *what* to label. We paid turkers a base amount for completing the tutorial and first mission ($0.82) and a bonus amount for each mission completed thereafter ($4.17/mile). These rates were based on US federal minimum wage ($7.25/hr), assuming an expected labeling rate of 1.74 miles/hr, which was drawn empirically from our data. In practice, our turkers earned $8.21/hr on average (*SD*=$5.99), which increased to $12.76 (*SD*=$6.60) for those 69 turkers who audited at least one mile. Turkers could see their earnings in real-time via the mission panel. We posted a total of 298 assignments over a 6-month period.

	Anonymous		Registered		Turkers		Researchers	
	All	Filtered	All	Filtered	All	Filtered	All	Filtered
Num users	384	293	243	188	170	122	28	21
% Filtered	--	23.7%	--	22.6%	--	28.2%	--	25.0%
Tot. miles	155.5	79.9	535.6	391.6	2,248.9	1,016.4	238.5	211.7
Avg. (SD)	0.4 (1.2)	0.3 (1.0)	2.2 (8.2)	2.1 (9.1)	13.2 (37)	8.3 (32)	8.5 (19)	10.1 (22)
Tot. missns	576	316	1,406	1,044	6,017	2,953	690	604
Avg (SD)	1.5 (3)	1.1 (2.5)	5.8 (20)	5.6 (22)	35.4 (95)	24.2 (87)	24.6 (53)	28.8 (62)
Tot. labels	12,950	10,760	41,588	35,923	150,847	103,820	31,954	30,488
Avg	33.7	36.7	171.1	191.1	887.3	851.0	1,141.2	1451.8
Lbls/100m	8.0	10.5	5.8	6.8	7.1	8.9	6.0	7.1
Avg. speed	1.22	0.74	1.93	1.58	1.68	1.14	2.76	2.57
Avg time	18.29	17.59	55.83	57.88	266.20	225.22	195.81	233.84
Avg desc	1.6	1.9	10.0	12.1	47.2	58.1	28.1	37.0

Table 2. The total amount of data collected during our deployment. Averages are per user. Avg. speed is in mi/hr, time is in mins, lbls/100m is median labels per 100m, and 'avg desc.' is # of open-ended descriptions.

Results

Overall, Project Sidewalk had 11,891 visitors to the landing page, of which 797 (627 volunteers; 170 turkers) completed the tutorial and audited at least one street segment in the first mission. In total, these users contributed 208,137 labels and audited 2,941 miles of DC streets (Table 1). Below, we analyze user behavior, contribution patterns, dropoff, and responses from a pop-up survey given to turkers. We examine worker and data quality in a separate section.

User behavior. On average, registered users completed more missions (5.8 *vs.* 1.5), contributed more labels (171.1 *vs.* 33.7), audited faster (1.93 mi/hr *vs.* 1.22), and spent more time on Project Sidewalk (55.8 mins *vs.* 18.3) than anonymous users (Table 2). Registered users also took longer on onboarding (6.9 mins *vs.* 3.8) and left more open-ended descriptions (10.0 *vs.* 1.6). Paid workers, however, did significantly more work on average than either volunteer group: 35.4 missions, 887.3 labels, and spent 4.4hrs using the tool. If we examine only those users who passed our "good" user heuristic, we filter 28.2% paid, 23.7% anonymous, and 22.6% registered workers; however, relative user behaviors stay the same. Similar to [29], user contribution patterns resembled a power law distribution: the top 10% anonymous, registered, and paid workers contributed 56.7%, 86.6%, and 80.2% of the labels in their group, respectively. By the top 25%, contributions rose to 77.4%, 93.6%, and 94.8%.

User dropoff. To examine user dropoff, we analyzed interaction logs for the last eight months of our deployment (after we added comprehensive logging to the tutorial). User dropoff was steep. While 1,110 users started the tutorial, only 568 finished it (51%), 479 (43.2%) took one step in their first mission, and 328 (29.5%) completed at least one mission. Of those 328, a majority, went on to finish their second mission (59.8%; 196 users) and then dropoff dampened substantially. For example, 74.0% of the users who completed *Mission 2* also completed *Mission 3*. When splitting the 1,110 users by group—846 volunteers and 264 turkers—we found different patterns of behavior. While only 43.9% of volunteers finished the tutorial and only 19.1% finished the first mission, turkers were far more persistent: 74.6% finished the tutorial and 62.9% completed the first mission.

Pop-up survey. To begin exploring why users contribute to Project Sidewalk, we developed a 5-question survey shown to users after their second mission. The first three questions asked about task enjoyment, difficulty, and self-perceptions of performance via 5-point Likert scales while the last two questions were open-ended asking about user motivation and feedback. A single researcher analyzed the two write-in questions via inductive coding. Though the survey is now given to all user groups, it was only available to turkers during our deployment study—which we analyze here.

In all, 123 turkers completed the survey. Of those, 110 (89.4%) stated that they enjoyed using Project Sidewalk (*M*=4.4; *SD*=0.7). For task difficulty, the responses were slightly more mixed: 83 turkers (67.5%) selected *easy* or *very easy* and 5 selected *difficult* (*avg*=3.9; *SD*=0.9). When asked to self-rate their performance, 81 turkers (65.9%) felt that they did at least a *very good* job and none reported *poor* (*avg*=4.0; *SD*=0.9). For the first open-ended question (required) about user motivation, 74 (60.2%) mentioned that the task was interesting or fun—"*It was an interesting and unique change to my day*" (U111); 48 (39.0%) felt that the task was important/helpful—"*I think it is important for those who are using wheelchairs to be able to safely navigate streets.*" (U223); and 20 (16.3%) mentioned money—"*It was interesting work and good pay*" (U61). The last question was optional and asked for feedback: 68 turkers chose to answer, mostly to thank us for the task (55 of 68): *"Good & interesting task. Thank you"* (U96). Six suggested features, five asked questions about labeling, and two reported bugs.

DATA VALIDATION STUDY

To investigate data quality and compare performance across user groups, we performed a data validation study using a subset of DC streets. This study occurred approximately halfway into our public deployment. Because pedestrian infrastructure can differ based on neighborhood type (*e.g.,* commercial *vs.* residential), age, and density, we first divided DC into four quadrants based on official geographic segmentation data [12]. We then sub-divided each quadrant into land-use zones using DC's open zoning regulation dataset [13]. Finally, we randomly selected the first two or three mission routes completed by individual volunteer users. This resulted in a test dataset of 44 miles (625 street segments) from 50 registered and 16 anonymous users across 62 of the 179 DC neighborhoods. We then verified that the selected routes had similar geographic and land-use distributions compared to all streets in DC.

To compare volunteer *vs.* paid worker performance, we posted the selected missions in our test dataset to Amazon Mechanical Turk. Other than payment, we attempted to carefully mimic the volunteer work experience: individual turkers completed onboarding and then were implicitly assigned either an *anonymous* user's mission set (two) or a *registered* user's mission set (three). To control for experience and learning effects, we did not allow deployment turkers to participate. We paid workers based on US federal minimum wage drawn from median volunteer completion times: \$2.00 for the tutorial + two missions (~2,000ft) and \$3.58 for the tutorial + three missions (~4,000ft). Unlike the deployment study, turkers could not choose to complete additional missions for bonus payment. To examine the effect of multiple labelers on performance, we hired five turkers per mission set for a total of 330 turkers.

To create ground truth, we first developed a labeling codebook based on ADA guidelines [51–53], which was then vetted and refined by a person who has used a wheelchair for 20 years. Following iterative coding [26], three researchers began labeling the same subset of data: one randomly selected mission set for an anonymous user and one for a registered user. For each round, the researchers met, resolved disagreements, and updated the codebook accordingly. After seven rounds, the average Krippendorff alpha score was 0.6 (*range*=0.5-0.8) and raw agreement: 85.4% (*SD*=4.1%). The three researchers then split the remaining 52 mission sets equally and a final review was performed. In total, ground truth consists of: 4,617 clusters, including 3,212 *curb ramps*, 1,023 *surface problems*, 295 *obstacles*, and 87 *missing curb ramps*. Though laborious, we note that the ground truth approach allows us to more deeply examine labeling performance compared with verifying placed labels—as the latter does not allow us to calculate false negatives.

Analysis. We examine accuracy at the street-segment level. We first cluster all labels from anonymous, registered, and paid workers using *single-user* clustering. We then use haversine distance to associate label clusters to their closest street segment. To compute our accuracy measures, we sum the number and type of label clusters for each segment and compare the result to ground truth. This produces counts of true/false positives and true/false negatives at each segment, which we binarize for final analysis. In total, 89.6% (560/625) of the street segments contained accessibility labels in ground truth. Unlike the four other label types, the *no sidewalk* label is not used for single-point targets but rather targets that extend multiple panoramas. Thus, we exclude this label from our analysis below.

We report on *raw accuracy* (number of segments that match ground truth), *recall*, and *precision*. Here, recall measures the fraction of accessibility targets that were found (labeled) compared to those in ground truth while precision measures the correctness of those labels. Ideally, each measure would be 1.0; however, similar to other crowdsourcing systems (*e.g.,* [25]), we prefer *high recall* over precision because correcting false positives is easier than false negatives—the former requires verification while the latter requires users actually re-explore an area. Except for the multiple labelers per segment analysis, we use only the *first* hired turker for each mission (rather than all five). For statistical analysis, we use binomial mixed effects models with user nested in mission route id and a logistic link function with accuracy, recall, and precision modeled as binomials. We assess significance with likelihood-ratio (LR) tests and use post-hoc

Tukey's HSD tests to determine statistical orderings. Our analysis was performed using the *R* statistical language.

Results

We examine overall performance across user groups, the effect of label type, label severity, and multiple labelers on accuracy, and common labeling mistakes.

User performance. The overall average raw accuracy was 71.7% (*SD*=13.0%) with all three user groups performing similarly (~70%). Because of the high true negative rates in our data—that is, most panoramas do *not* have accessibility issues and were correctly labeled that way—recall and precision are more insightful measures (Figure 5). Turkers found significantly more issues compared with registered and anonymous (*recall*=67.8% *vs.* 61.4% *vs.* 48.8%, respectively) at similar precision levels (68.8% *vs.* 72.2% vs. 74.5%). Using an LR test, we found *user group* had a statistically significant association with recall (lr=21.6, df=2, n=132, p<0.001) and precision (lr=7.1, df=2, n=131, p=0.028) but not raw accuracy. Pairwise comparisons for recall were all significant but none were for precision.

To explore the effect of multiple labelers on performance, recall that we hired five turkers per mission set. We examine *majority vote* for each group size (3, 5) as well as treating each contribution individually (*e.g.,* Turk3$_{maj}$ *vs.* Turk3$_{all}$). We expect that Turk$_{maj}$ will result in higher precision but lower recall as it requires more than one user to label the same target and just the opposite from Turk$_{all}$ (*i.e.,* higher recall, lower precision). Indeed, this is what we found: from Turk1 (baseline) to Turk5$_{all}$, recall rose from 67.8% to 91.7% but at a cost of precision (from 68.8% to 55.0%). In contrast, for majority vote, recall fell from 67.8% to 59.5% but precision rose from 68.8% to 87.4%. Again, we found *turker group* had a statistically significant association with recall (lr=498.96, df=4, n=330, p<0.001) and precision (lr= 374.88, df=4, n=330, p<0.001). All pairwise comparisons for recall and precision were significant except for Turk5$_{maj}$ < Turk3$_{maj}$—for recall only.

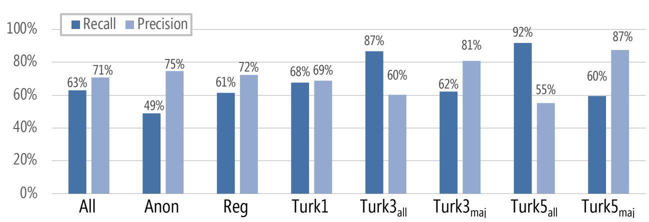

Figure 5. Average recall and precision for all user groups.

Label type. To examine accuracy as a function of label type, we analyzed labeling data across users (Table 3). *Curb ramps* were the most reliably found and correctly labeled with *recall*=86.0% and *precision*=95.4% respectively. In contrast, while *no curb ramps* had reasonably high recall at 69.3%, precision was only 20.5% suggesting an incorrect understanding of what justifies a *no curb ramp* label. The other two label types, *obstacle* and *surface problem*, had lower recall (39.9% and 27.1%) but comparatively higher precision (47.5% and 72.6%), which mirrors our experience with ground truth—these accessibility problems are hard to find and require diligent exploration. In addition, these two label types can legitimately be switched in some cases (*e.g.,* a patch of overgrown grass could be marked as either an *obstacle* or *surface problem*). We explore labeling mistakes in more detail below.

	Gnd Truth Clusters	Raw Acc.	Recall	Precision
Curb Ramp	3,212	83.7 (23.1)	86.0 (25.7)	95.4 (7.5)
No Curb Ramp	87	72.9 (21.9)	69.3 (43.5)	20.5 (31.7)
Obstacle	295	71.2 (18.8)	39.9 (36.9)	47.5 (37.4)
Surface Problem	1,023	59.0 (24.8)	27.1 (30.5)	72.6 (35.4)

Table 3. Accuracy by label type. All pairwise comparisons are significant.

Effect of severity. We hypothesized that high-severity problems would be easier to find. To explore this, we partitioned ground truth labels into two groups: *low severity* (<= 2 rating) and *high severity* (>= 3 rating). The low severity group contained 1,053 labels and the high 352 labels. As expected, we found that high-severity labels had significantly higher recall (*Mdn*=83.3%; *avg*=69.8%; *SD*=35.5%) than low-severity labels (*Mdn*=56.3%; 57.0%; *SD*=32.3%). To determine significance, we created a binomial mixed effect model with *severity* (high or low) as the fixed effect and *user* nested in *mission route id* as random effects. Result of LR test (lr=10.6, df=1, n=246, p=0.001).

Common Labeling Errors

To better understand labeling errors and to contextualize our quantitative findings, we conducted a qualitative analysis of labeling errors. We randomly selected 54 false positives and 54 false negatives for each label type, which resulted in 432 total error samples from 127 anonymous, 141 registered, and 210 paid workers. A single researcher inductively analyzed the data with an iteratively created codebook. We show the top three errors with examples in Figure 6. In analyzing false positives, we observed that most mistakes were understandable and either easy to correct with better training or innocuous. For example, 66.7% of incorrect *curb ramp* labels were applied to driveways, nearly half of *obstacles* and *surface problems* were potentially legitimate issues but not on the primary pedestrian route (*e.g.,* middle of street *vs.* crosswalk), and almost 30% of incorrect *missing curb ramps* were on extended residential walkways. Moreover, 32.7% of *surface problems* and 9.3% of *obstacles* were correctly labeled as problems but with a different label type from ground truth—*e.g.,* a surface problem marked as an obstacle.

For false negatives (*i.e.,* a user did not label a problem when one exists), it is harder to discern clear patterns—at least for some label types. For *obstacles* and *surface problems*—both of which had the lowest recall and thus can be considered hardest to find—salience appears to be a contributing factor: 50% of missed *obstacles* were only partially blocking the pedestrian path and nearly 30% of *surface problems* were grass-related. For *missing curb ramps*, 46.3% of missed labels were at a corner where at least one other curb ramp exists though the second most common error was more egregious: a pedestrian path to a street had no curb ramp and

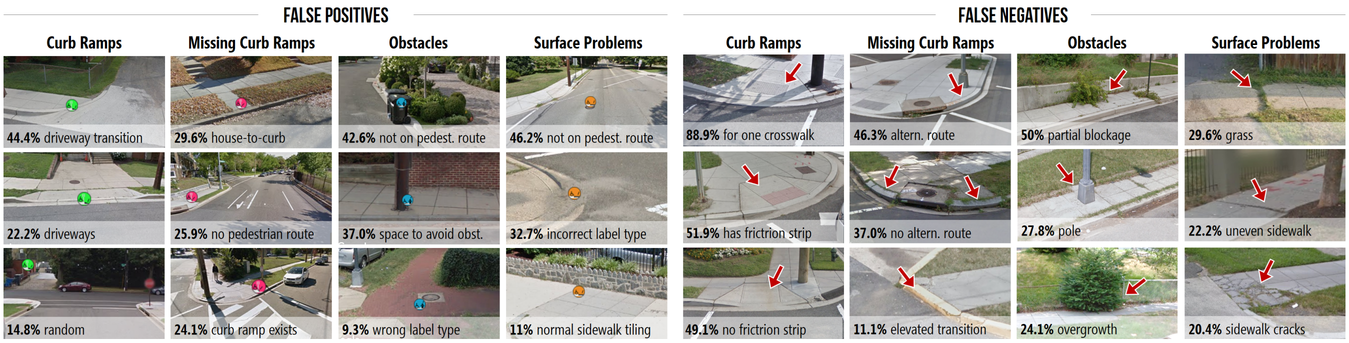

Figure 6. An overview of false positive and negative labeling mistakes ordered by frequency (taken from 432 error samples in the data validation study).

no alternative accessible route (37.0%). We discuss potential solutions to address labeling errors in the Discussion.

SEMI-STRUCTURED INTERVIEW STUDY

To complement our deployment and data validation studies and to solicit reactions to Project Sidewalk from key stakeholders, we conducted an interview study with three DC-area groups (*N*=14): six *government officials* (G), five *people with mobility impairments* (MI), and three *caregivers* (C). G included state and city transportation employees with oversight of pedestrian infrastructure, MI participants used a mobility aid such as a wheelchair or cane, and caregivers took care of a person with a MI either as a professional, family member, or friend. Participants were recruited via mailing lists, word-of-mouth, and social media.

The three-part study began with a semi-structured interview about participants' current perceptions of and problems with urban accessibility. We then asked participants to use Project Sidewalk while "thinking aloud." Finally, we concluded with a debrief interview about the tool, including its perceived utility, concerns, and design ideas. Sessions lasted between 60-65 minutes, and participants were compensated $25. One government session was a group interview with three participants (coded G3); all other interviews were individual. Sessions were audio- and screen-recorded, which were transcribed and coded to find emergent themes using peer debriefing [10,48]. Using deductive coding, one researcher created an initial codebook for the interviews, which was refined with the help of a peer. A randomly selected transcript was then coded, which was reviewed by a second researcher using peer-debriefing. To resolve conflicts and update the codebook, the two researchers met after each review process. The final codebook was produced after three iterations (one transcript per stakeholder group). The remaining data was then coded by the initial researcher.

Results

We describe findings related to the perceived value and usability of Project Sidewalk as well as design suggestions and concerns. For quotes, we use (participant group + id).

Perceived value. Overall, all three stakeholder groups felt that Project Sidewalk enabled rapid data collection, allowed for gathering diverse perspectives about accessibility, and helped engage citizens in thinking about urban design. Government officials emphasized cost-savings and community involvement envisioning Project Sidewalk as a triaging tool before sending out employees to physically examine areas: "*It's really good for a starting point. This is a first observation, and when you send somebody out in the field, they can see those observations and pick up more information. It's just neat*" (G4). The MI and caregiver groups focused more on personal utility, envisioning accessibility-aware navigation tools that could incorporate Project Sidewalk data: "*I might take advantage of more opportunities knowing that, okay, if I could rely on the data and knew I could anticipate how difficult it was going to be for me to get around*" (MI1). Six of the seven MI and caregiver participants mentioned that Project Sidewalk data could enhance their independence, give them confidence to explore new and unfamiliar areas, and/or help them achieve the same pedestrian rights as everyone else.

Usability. Participants across groups felt that the tool was easy-to-learn and fun to use. G3, for example, stated: "*I think it's awesome. [...] It's a lot of fun*" and reported "*feeling good*" contributing data to a social purpose while also being motivated by the game design elements: "*we're looking at the 71 percent complete, and we're pretty excited!*" Three participants appreciated relying on a familiar technology like GSV, "*You're not introducing like yet another platform that somebody has to relearn—that was helpful*" (G3). Almost everyone (13/14) found the labeling system comprehensive as captured by MI3: "*the labeling is pretty all-inclusive.*"

Concerns. Key concerns included outdated GSV imagery or labels (*N*=6), data reliability (3), and conflicting data (4). Towards outdated imagery and labels, C1 asked "*if a street light was marked as an obstacle and if it was replaced or moved, would the labels reflect that?*" While this is one limitation of our virtual auditing approach, four participants mentioned that they would rather be aware of a potential issue even if it no longer existed. For example, C2 stated: "*if there was a label, I'd rather be aware of it.*" For data reliability, G4 suggested that each road be audited by multiple people: "*I would have more confidence if different people did it, did the same street.*" Four participants (2 Cs, 2 MIs) were concerned about how labelers may differ in interpreting problems compared with their needs and experiences. For example, MI1 said: "*my concern as a user…someone said this was accessible and I got there and* *it wasn't accessible, because everyone has different opinions on accessibility.*"

Suggestions. Participants suggested developing mechanisms to keep information up-to-date (4)—for example, by adding a complementary smartphone-based data collection app, adding verification interfaces (3), and surfacing data age (2). All government officials were interested in ways to export and visualize the data; one suggested integrating directly into their service request backend. At a more detailed tool level, seven participants suggested adding new label types, including for crosswalks, the presence of sidewalks, access points (such as driveways), and construction.

DISCUSSION AND CONCLUSION

Through a multi-methods approach, our results demonstrate the viability of virtually auditing urban accessibility at scale, highlight behavioral and labeling quality differences between user groups, and summarize how key stakeholders feel about Project Sidewalk and the crowdsourced data. Below, we discuss worker and data quality, future deployment costs and worker sources, and limitations.

Label quality. Our data validation study found that, on average, users could find 63% of accessibility issues at 71% precision. This is comparable to early streetscape labeling work by Hara *et al.* [24], where turkers labeled at 67.0% and 55.6% for recall and precision, respectively; however, our tasks are more complex, contain more label types, and are evaluated at a larger scale. Like [24], we also show how assigning multiple labelers can improve results and describe tradeoffs in aggregation algorithms—*e.g.,* by combining labels from five turkers per street, recall rose to 92%; however, precision fell from 69% to 55%. We believe our findings represent a lower bound on performance and provide a nice baseline for future work.

To improve quality, we envision four areas of future work: first, a more sophisticated workflow pipeline that dynamically verifies labels [6,45], allocates the number of assigned labelers/street based on inferred performance, and integrates other datasets (*e.g.,* top-down imagery). Second, though not explored in this paper, our mission-based architecture supports a large variety of diverse mission tasks—*e.g.,* verification missions and ground truth seeding missions, both which will enable us to more reliably identify poor-quality workers. Third, Project Sidewalk currently relies solely on manual labeling; we are experimenting with deep learning methods trained on our 240,000+ image-based label dataset to detect problems automatically (building on [25,49]), triage likely problem areas, and/or aid in verifications. Finally, our results suggest that many *false positives* could be corrected via improved training (*e.g.,* a driveway is not a curb ramp) and by using simple automated validation (*e.g.,* check for labels in unlikely areas).

Data age. Our interview participants raised two concerns about data age: GSV image age and label age. Towards the former, prior work has found high agreement between virtual audit data of pedestrian infrastructure compared with traditional audits [5,9,22,25,46,56] . Google does not publish how often their GSV cars collect data; however, in a 2013 analysis of 1,086 panorama sampled across four North American cities, the average age was 2.2yrs (SD=1.3) [25]. In our dataset, workers labeled 74,231 panoramas, which at the time of first label, were also M=2.2yrs old (SD=1.5). As a comparison, the official opendata.dc.gov curb ramp dataset [14] was captured in 1999 and last updated in 2010 (no other label types are included). Moreover, our general approach should work with any streetscape imagery dataset, including *Mapillary* [29], *CycloMedia,* or *Bing StreetSide*—many of which are exploring high-refresh methods via automated vehicles and crowd contributions. In terms of maintaining labels over time, one benefit of our scalable approach is that streets can be periodically re-audited and old labels can be used to study historical change (*e.g.,* [38]).

Cost. While future deployments could rely solely on paid workers, ideally Project Sidewalk would also engage online and local communities who are concerned with urban accessibility. Based on our deployment study, we estimate that auditing DC with 100 paid workers alone would cost $34,000 and take 8 days (assuming five labelers/street, 8hrs of work per day, and that 72 of 100 met our "good" user quality threshold). If one-third of DC was audited by volunteers, costs fall below $25,000. DC is a large city and has a well-resourced DOT with fulltime ADA compliance staff; small-to-medium sized cities often lack ADA budgets and could particularly benefit from Project Sidewalk. Indeed, we have been contacted by more than a dozen cities in the US and Canada about future deployments.

Limitations. There are three main limitations with crowdsourcing virtual audits: panorama age, label quality, and the ability for crowdworkers to see and assess sidewalks from GSV. We addressed the former two points above. Towards the latter, users could mark areas as occluded in our tool (*e.g.,* a truck blocking a sidewalk); however, *occlusion* constituted only 0.4% of all applied labels in our deployment suggesting that most sidewalks are visible. For study limitations, we employed a multi-methods approach to mitigate the effects of any one study technique. Still, longitudinal deployment studies are messy and ours is no exception: we lost over two months of deployment time due to changes in the GSV API, maintenance upgrades to our servers, and personnel changes. For the data validation study, we were unable to consistently reach high α agreement for *obstacles* and *surface problems* during our seven iterative rounds of coding; these label types are challenging and can be legitimately conflated (*e.g.,* marking overgrown grass as a surface problem *vs.* an obstacle). Our performance results for these label types may have been impacted. Finally, while our studies take place in the US, accessible infrastructure is a global problem. Project Sidewalk should work wherever there is GSV and OSM.

REFERENCES

1. Court of Appeals 3rd Circuit. 1993. *Kinney v. Yerusalim, 1993 No. 93-1168*. Retrieved from http://www.access-board.gov/prowac/yerusalim.htm

2. Ahmed Loai Ali, Nuttha Sirilertworakul, Alexander Zipf, Amin Mobasheri, Ahmed Loai Ali, Nuttha Sirilertworakul, Alexander Zipf, and Amin Mobasheri. 2016. Guided Classification System for Conceptual Overlapping Classes in OpenStreetMap. *ISPRS International Journal of Geo-Information* 5, 6: 87. https://doi.org/10.3390/ijgi5060087

3. V Antoniou and A Skopeliti. 2015. Measures and Indicators of VGI Quality: An Overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* II-3/W5: 345–351. Retrieved from https://search.proquest.com/docview/1987376210?accountid=14784

4. Michael D.M. Bader, Stephen J. Mooney, Yeon Jin Lee, Daniel Sheehan, Kathryn M. Neckerman, Andrew G. Rundle, and Julien O. Teitler. 2015. Development and deployment of the Computer Assisted Neighborhood Visual Assessment System (CANVAS) to measure health-related neighborhood conditions. *Health & Place* 31: 163–172. https://doi.org/10.1016/J.HEALTHPLACE.2014.10.012

5. Hannah M Badland, Simon Opit, Karen Witten, Robin A Kearns, and Suzanne Mavoa. 2010. Can virtual streetscape audits reliably replace physical streetscape audits? *Journal of urban health : bulletin of the New York Academy of Medicine* 87, 6: 1007–16. https://doi.org/10.1007/s11524-010-9505-x

6. Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 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

7. John R Bethlehem, Joreintje D Mackenbach, Maher Ben-Rebah, Sofie Compernolle, Ketevan Glonti, Helga Bárdos, Harry R Rutter, Hélène Charreire, Jean-Michel Oppert, Johannes Brug, and Jeroen Lakerveld. 2014. The SPOTLIGHT virtual audit tool: a valid and reliable tool to assess obesogenic characteristics of the built environment. *International Journal of Health Geographics* 13, 1: 52. https://doi.org/10.1186/1476-072X-13-52

8. Carlos Cardonha, Diego Gallo, Priscilla Avegliano, Ricardo Herrmann, Fernando Koch, and Sergio Borger. 2013. A Crowdsourcing Platform for the Construction of Accessibility Maps. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (W4A '13), 26:1--26:4. https://doi.org/10.1145/2461121.2461129

9. Philippa Clarke, Jennifer Ailshire, Robert Melendez, Michael Bader, and Jeffrey Morenoff. 2010. Using Google Earth to conduct a neighborhood audit: Reliability of a virtual audit instrument. *Health & Place* 16, 6: 1224–1229. https://doi.org/10.1016/J.HEALTHPLACE.2010.08.007

10. John W. Creswell. 2012. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*. Sage Publications, Inc.

11. Igor Gomes Cruz and Claudio Campelo. 2017. Improving Accessibility Through VGI and Crowdsourcing. In *Volunteered Geographic Information and the Future of Geospatial Data*, Claudio Campelo, Michela Bertolotto and Padraig Corcoran (eds.). 208–226.

12. DC.gov. OpenData DC Quadrants. Retrieved September 1, 2018 from http://opendata.dc.gov/datasets/02923e4697804406b9ee3268a160db99_11

13. DC.gov. OpenData DC Zoning Regulations of 2016. Retrieved September 1, 2018 from http://opendata.dc.gov/datasets/zoning-regulations-of-2016

14. DC.gov. OpenData DC Sidewalk Ramp Dataset. Retrieved September 1, 2018 from http://opendata.dc.gov/datasets/sidewalk-ramps-2010

15. Chaohai Ding, Mike Wald, and Gary Wills. 2014. A Survey of Open Accessibility Data. In *Proceedings of the 11th Web for All Conference* (W4A '14), 37:1--37:4. https://doi.org/10.1145/2596695.2596708

16. Google Inc. 2018. Google Street View Service. Retrieved September 1, 2018 from https://developers.google.com/maps/documentation/javascript/streetview

17. David Gutman. Seattle may have to spend millions making sidewalks more accessible to people with disabilities | The Seattle Times. *The Seattle Times*. Retrieved September 1, 2018 from https://www.seattletimes.com/seattle-news/transportation/seattle-may-have-to-spend-millions-making-sidewalks-more-accessible/

18. Richard Guy and Khai Truong. 2012. CrossingGuard: exploring information content in navigation aids for visually impaired pedestrians. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)* (CHI '12), 405–414. https://doi.org/10.1145/2207676.2207733

19. Mordechai Haklay. 2010. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets. *Environment and Planning B: Planning and Design* 37, 4: 682–703. https://doi.org/10.1068/b35097

20. Mordechai (Muki) Haklay, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. 2010. How Many Volunteers Does it Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal* 47, 4: 315–322. https://doi.org/10.1179/000870410X12911304958827

21. Muki Haklay. 2013. Citizen Science and Volunteered

Geographic Information: Overview and Typology of Participation. In *Crowdsourcing Geographic Knowledge*. Springer Netherlands, Dordrecht, 105–122. https://doi.org/10.1007/978-94-007-4587-2_7

22. Kotaro Hara, Shiri Azenkot, Megan Campbell, Cynthia L Bennett, Vicki Le, Sean Pannella, Robert Moore, Kelly Minckler, Rochelle H Ng, and Jon E Froehlich. 2015. Improving Public Transit Accessibility for Blind Riders by Crowdsourcing Bus Stop Landmark Locations with Google Street View: An Extended Analysis. *ACM Transactions on Accessible Computing (TACCESS)* 6, 2: 5:1--5:23. https://doi.org/10.1145/2717513
23. Kotaro Hara, Christine Chan, and Jon E Froehlich. 2016. The Design of Assistive Location-based Technologies for People with Ambulatory Disabilities: A Formative Study. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 1757–1768. https://doi.org/10.1145/2858036.2858315
24. Kotaro Hara, Vicki Le, and Jon Froehlich. 2013. Combining crowdsourcing and google street view to identify street-level accessibility problems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*, 631–640. https://doi.org/10.1145/2470654.2470744
25. Kotaro Hara, Jin Sun, David W Jacobs, and Jon E Froehlich. 2014. Tohme: Detecting Curb Ramps in Google Street View Using Crowdsourcing, Computer Vision, and Machine Learning. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, TBD.
26. Daniel J. Hruschka, Deborah Schwartz, Daphne Cobb St.John, Erin Picone-Decaro, Richard A. Jenkins, and James W. Carey. 2004. Reliability in Coding Open-Ended Data: Lessons Learned from HIV Behavioral Research. *Field Methods* 16, 3: 307–331. https://doi.org/10.1177/1525822X04266540
27. Winnie Hu. For the Disabled, New York's Sidewalks Are an Obstacle Course - The New York Times. *New York Times*. Retrieved September 1, 2018 from https://www.nytimes.com/2017/10/08/nyregion/new-york-city-sidewalks-disabled-curb-ramps.html
28. Yusuke Iwasawa, Kouya Nagamine, Ikuko Eguchi Yairi, and Yutaka Matsuo. 2015. Toward an Automatic Road Accessibility Information Collecting and Sharing Based on Human Behavior Sensing Technologies of Wheelchair Users. *Procedia Computer Science* 63: 74–81. https://doi.org/10.1016/J.PROCS.2015.08.314
29. Levente Juhász and Hartwig H. Hochmair. 2016. User Contribution Patterns and Completeness Evaluation of Mapillary, a Crowdsourced Street Level Photo Service. *Transactions in GIS* 20, 6: 925–947. https://doi.org/10.1111/tgis.12190
30. Reuben Kirkham, Romeo Ebassa, Kyle Montague, Kellie Morrissey, Vasilis Vlachokyriakos, Sebastian Weise, and Patrick Olivier. 2017. WheelieMap: An Exploratory System for Qualitative Reports of Inaccessibility in the Built Environment. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (MobileHCI '17), 38:1--38:12. https://doi.org/10.1145/3098279.3098527
31. Anthony Li, Manaswi Saha, Anupam Gupta, and Jon E. Froehlich. 2018. Interactively Modeling And Visualizing Neighborhood Accessibility At Scale: An Initial Study Of Washington DC. In *Poster Proceedings of ASSETS'18*.
32. Afra Mashhadi, Giovanni Quattrone, and Licia Capra. 2013. Putting Ubiquitous Crowd-sourcing into Context. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (CSCW '13), 611–622. https://doi.org/10.1145/2441776.2441845
33. Afra Mashhadi, Giovanni Quattrone, Licia Capra, and Peter Mooney. 2012. On the Accuracy of Urban Crowd-sourcing for Maintaining Large-scale Geospatial Databases. In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration* (WikiSym '12), 15:1--15:10. https://doi.org/10.1145/2462932.2462952
34. Hugh Matthews, Linda Beale, Phil Picton, and David Briggs. 2003. Modelling Access with GIS in Urban Systems (MAGUS): capturing the experiences of wheelchair users. *Area* 35, 1: 34–45. https://doi.org/10.1111/1475-4762.00108
35. Andrew May, Christopher J. Parker, Neil Taylor, and Tracy Ross. 2014. Evaluating a concept design of a crowd-sourced 'mashup' providing ease-of-access information for people with limited mobility. *Transportation Research Part C: Emerging Technologies* 49: 103–113. https://doi.org/10.1016/J.TRC.2014.10.007
36. Allan R Meyers, Jennifer J Anderson, Donald R Miller, Kathy Shipp, and Helen Hoenig. 2002. Barriers, facilitators, and access for wheelchair users: sbstantive and methodologic lessons from a pilot study of environmental effects. *Social Science & Medicine* 55, 8: 1435–1446. https://doi.org/10.1016/S0277-9536(01)00269-6
37. Amin Mobasheri, Jonas Deister, and Holger Dieterich. 2017. Wheelmap: the wheelchair accessibility crowdsourcing platform. *Open Geospatial Data, Software and Standards* 2, 1: 27. https://doi.org/10.1186/s40965-017-0040-5
38. Ladan Najafizadeh and Jon E. Froehlich. 2018. A Feasibility Study of Using Google Street View and Computer Vision to Track the Evolution of Urban Accessibility. In *Poster Proceedings of ASSETS'18*.
39. Andrea Nuernberger. 2008. Presenting Accessibility to Mobility-Impaired Travelers (Ph.D. Thesis). University of California, Santa Barbara.
40. OpenSidewalks.com. OpenSidewalks. Retrieved September 1, 2018 from https://www.opensidewalks.com/
41. Katherine Panciera, Reid Priedhorsky, Thomas Erickson, and Loren Terveen. 2010. Lurking? cyclopaths?: a

quantitative lifecycle analysis of user behavior in a geowiki. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 1917–1926. https://doi.org/10.1145/1753326.1753615

42. Falko Weigert Petersen, Line Ebdrup Thomsen, Pejman Mirza-Babaei, and Anders Drachen. 2017. Evaluating the Onboarding Phase of Free-toPlay Mobile Games: A Mixed-Method Approach. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (CHI PLAY '17), 377–388. https://doi.org/10.1145/3116595.3125499

43. Catia Prandi, Paola Salomoni, and Silvia Mirri. 2014. mPASS: Integrating people sensing and crowdsourcing to map urban accessibility. In *2014 IEEE 11th Consumer Communications and Networking Conference (CCNC)*, 591–595. https://doi.org/10.1109/CCNC.2014.6940491

44. Giovanni Quattrone, Afra Mashhadi, Daniele Quercia, Chris Smith-Clarke, and Licia Capra. 2014. Modelling Growth of Urban Crowd-sourced Information. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining* (WSDM '14), 563–572. https://doi.org/10.1145/2556195.2556244

45. Alexander J Quinn and Benjamin B Bederson. 2011. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '11), 1403–1412. https://doi.org/10.1145/1978942.1979148

46. Andrew G. Rundle, Michael D.M. Bader, Catherine A. Richards, Kathryn M. Neckerman, and Julien O. Teitler. 2011. Using Google Street View to Audit Neighborhood Environments. *American Journal of Preventive Medicine* 40, 1: 94–100. https://doi.org/10.1016/J.AMEPRE.2010.09.034

47. Daniel Sinkonde, Leonard Mselle, Nima Shidende, Sara Comai, Matteo Matteucci, Daniel Sinkonde, Leonard Mselle, Nima Shidende, Sara Comai, and Matteo Matteucci. 2018. Developing an Intelligent PostGIS Database to Support Accessibility Tools for Urban Pedestrians. *Urban Science* 2, 3: 52. https://doi.org/10.3390/urbansci2030052

48. Sharon Spall. 1998. Peer Debriefing in Qualitative Research: Emerging Operational Models. *Qualitative Inquiry* 4, 2: 280–292. https://doi.org/10.1177/107780049800400208

49. Jin Sun and David W. Jacobs. 2017. Seeing What is Not There: Learning Context to Determine Where Objects are Missing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1234–1242. https://doi.org/10.1109/CVPR.2017.136

50. U.S. Census Bureau. U.S. Census QuickFacts: District of Columbia. Retrieved September 1, 2018 from https://www.census.gov/quickfacts/fact/table/dc/PST045217

51. U.S. Department of Transportation Federal Highway Administration. A Guide for Maintaining Pedestrian Facilities for Enhanced Safety. Retrieved September 1, 2018 from https://safety.fhwa.dot.gov/ped_bike/tools_solve/fhwasa13037/chap3.cfm

52. United States Access Board. Accessible Guidelines and Standards for Public Rights-of-Way. Retrieved from https://www.access-board.gov/guidelines-and-standards/streets-sidewalks/public-rights-of-way/proposed-rights-of-way-guidelines/chapter-r3-technical-requirements

53. United States Department of Justice. 2010. 2010 ADA Standards for Accessible Design. Retrieved September 1, 2018 from https://www.ada.gov/regs2010/2010ADAStandards/2010ADAstandards.htm

54. United States Department of Justice Civil Rights Division. 1990. *Americans with Disabilities Act of 1990, Pub. L. No. 101-336, 104 Stat. 328*.

55. Washington.org. Washington DC Visitor Research. Retrieved September 1, 2018 from https://washington.org/press/dc-information/washington-dc-visitor-research

56. Jeffrey S. Wilson, Cheryl M. Kelly, Mario Schootman, Elizabeth A. Baker, Aniruddha Banerjee, Morgan Clennin, and Douglas K. Miller. 2012. Assessing the Built Environment Using Omnidirectional Imagery. *American Journal of Preventive Medicine* 42, 2: 193–199. https://doi.org/10.1016/J.AMEPRE.2011.09.029

57. J.O. Wobbrock, S.K. Kane, K.Z. Gajos, S. Harada, and J. Froehlich. 2011. Ability-Based Design: Concept, Principles and Examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3: 9. Retrieved June 16, 2011 from http://portal.acm.org/citation.cfm?id=1952384