How Geotagged Social Media Can Inform Modern Travelers

Dan Tasse
Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
dan.tasse@gmail.com

CMU-HCII-17-102
May 2017
Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Committee:
Jason Hong, Carnegie Mellon University, Chair
Jodi Forlizzi, Carnegie Mellon University
Aniket Kittur, Carnegie Mellon University
Judd Antin, Airbnb

Keywords: Social media, creative tourism, geotags

ABSTRACT

Modern tourists travel in new ways. The rising class of so-called "Creative tourists" prefer to explore everyday life instead of simply ticking off a list of sights to see. However, travel guides all currently represent places as simply a collection of sights.

At the same time, public geotagged social media data is opening a new world of ways to investigate another place. In this thesis, I describe efforts to bring these trends together, by developing neighborhood guides for travelers, based on social media. I first investigate why people geotag and where this public geotagged data comes from. Then, after developing a model of what tourists want through a series of interviews and surveys, I develop a prototype social-media-based neighborhood guide for travelers. By an iterative user study and quantitative investigation into photo sources, I find that this data can give users an ideal glimpse into a new city.

Implications are widespread: I show not only how social media can be used to help people travel, but also develop a perspective on what social media tells, and does not tell, about cities and neighborhoods. I show that social media provides an idealized qualitative image into a city, while perhaps not reflecting the objective, quantitative reality. This matches tourists' needs ideally, providing an exciting new opportunity for a new generation of tourism tools.

ACKNOWLEDGEMENTS

0.1 INTRODUCTION

Man, thank you to so many people. This has probably been the hardest thing I've ever done, which I guess says something about growing up an upper-middle-class white man in the US, but regardless, it was still hard, and I certainly wouldn't have gotten through this on my own.

0.2 RELATED PEOPLE

First, we've got to do the easy ones. Thank you to Tati for being incredibly supportive and just the absolute best, even though this whole PhD process *must* have constantly brought out the worst in me. And thank you to CMU for unwittingly bringing us together. If I had spent these 5 years doing nothing else but meeting you, Tati, it would be 100% worth it.

My family too, of course. Thanks, Mom, Dad, and Cheryl, for always being there for me, PhD world and beyond. For funding my way through CMU the first time so I could come back on my own dime. And to my family that has become my family over these last 5 years: Jon, and Dana, Lukas, Natasha, and Ivana.

0.3 METHODICALLY GREAT PEOPLE

I guess next most important in getting me to graduate are the people who logistically made it happen: all the professors who took a chance on me. Chronologically, I've got to start with Noah Smith, who made time to meet with me, a confused undergrad who didn't know what the academic world meant, stuck with me when my undergrad project ended up kind of a mess, and still was happy to help me get back into the academic world. Shwetak Patel, who allowed me to just start showing up in his lab in 2011 and working on stuff, and Julie Kientz, who let me run an entire project that ended up trying to bite off way more than I could chew, but still wrote me a nice letter to convince these folks at CMU to accept me.

Thank you to Anind Dey, for being willing to take me on even though our academic interest overlap was really just "ubiquitous computing is cool," and letting me realize myself that that's not really a research direction. Joshua Hailpern and Anupriya Ankolekar for taking me on as an intern, despite that our project fizzled out thoroughly by internship day 5, and then Ayman Shamma and Bart Thomee for taking yet another chance on me at really a pretty critical time. Thank you to Ayman in particular; I don't know if I'd have finished this whole thing without your support. (Special shout-out to the HCI Research team, and then the Vision team, for being my home away from home while I was finishing up. I really appreciate how much you welcomed me in. Thesisizing is lonely and having some friends really helps.)

Thank you to my committee, Niki and Jodi and Judd. I know, it's maybe not the most intensive

working relationship you've had, but I appreciate knowing that you're in my corner.

And thank you to Jason. Thanks for advising me when I was super lost here, thanks for continuing to believe in me when I cooked up not only a bad pancake but a whole bad breakfast, thanks for spending probably all your unrestricted grants to fund me, thanks for just really being kind and easy to work with. I hate to squash this paragraph in the middle here; I feel like it should be dramatically at the beginning or the end of a section. I will say, if there's one person who explained most of the variance of me finishing this thing, it's got to be Jason, so thanks for that.

Coworkers and coauthors! Thanks to Alex, Zichen, JZ, Will, Jason W, Beka, Jenny, and Dave; in some cases you gave me a little boost, in some cases you carried me on your back. In either case, I appreciate it. And thanks to Himanshu, Erick, Andy, Jennifer, Hong Bin, Emily, Josh, Jinny, Eva, and Sriram for putting up with me as I learn how to manage or mentor people. I hope I didn't steer you too far astray.

Before we leave this section, I want to also thank Lauren, Ja'Ron, Justin, and Jessica for personally helping me get things done here. (or sometimes, doing the things themselves.) And of course, Queenie. Wouldn't have been able to do it without you either.

0.4 Results: Lasting Friendships

Thank you to my wonderful cohort: Chris, Caitlin, Nikola, Annie, Dave, Vivian, Jenny, Ryan, Brandon, JoAnna, Anthony, and Shuang. I didn't know at first why they made such a big deal at the HCII about this, but realized that it's really hard to get through it if you don't have others you can commiserate with. Thank you to my amazing puzzling squad: Kelly, Erik, Robert, and Julia. Tati and the rest of the obscure restaurants clique: Jeff, Beka, Sauvik, and Anna. Album Club, which I think is a subset of these people. And, I mean, a blanket thank you to all the CMU friends I haven't mentioned here; there isn't a single HCII student that I don't wish I could have spent more time with.

I picked CMU for grad school primarily because I got along so well with the students. That turned out to be absolutely correct.

Thanks also to my outside-CMU friends, who reminded me that there *is* another world out there; that was really important, especially in the most difficult times. Thank you to Beej, Nicole, Aaron, Mel, Greg, Ruth, Nils, Anne, and all my kickball teammates along the way. And thank you to amazing housemates and friends Emmy and Sethie, and honorary housemate Farmer Jim.

0.5 Discussion: Mental Health is Important

Not to get too heavy before even the Table of Contents, but the depressed grad student is a trope for a reason. Grad school is hard. It might wear on your mental health. If it is, you should really get that help.

Thank you to Michael at CAPS for starting the ball rolling on me getting help at all. Thank you to the first other therapist I started seeing, and my psychiatrist out here, whose names I forget; thank you to Joshua out here in SF for keeping me sane through the remote-working

home stretch. And thank you to Ryan; I wish you could have had someone who'd helped you as much as you'd helped me.

0.6 CONCLUSION

Thank you for everything; I have no complaints whatsoever.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

People travel differently than they used to. They want to "understand the feel of an area", to "see everyday life", to "live like a local" instead of just "seeing the tourist sights." Part of this is due to the excesses of the tourism industry, part of it is due to the ever-decreasing cost of leisure travel, and part of it is due to increasing curiosity about how our fellow humans live. The shift is large enough that some have called it a demographic shift, describing modern travelers as "creative tourists" [76], "new urban tourists" [23], or "Explorers" [93]. Airbnb, for example, suggests Explorers may be 43% of their user base, numbering in the tens of millions [93]. But in an unfamiliar place, these creative tourists may wonder where to go. After all, some parts of any city may be boring, dangerous, or otherwise unsuitable. Travelers need to be able to understand the neighborhoods of a city to plan out an enjoyable trip.

This is a specific case of a general problem: we need new ways to understand cities and neighborhoods. As more people move to cities throughout the 21st century, quickly understanding how places feel will become more and more important. People moving will need to know what neighborhood they would feel at home in, business owners will need to know where to expand and market, and city planners will need to know how to allocate services and zone districts.

Travelers have a unique set of information needs, though, because they are new to a place and do not have time to build up local knowledge from experience. Unlike the sun-and-sand tourists of two generations ago or the cultural-site-visiting tourists of last generation [19], today's tourists want to curate and create their own experience [76]. And more than ever, platforms like Airbnb and Couchsurfing help them do so by staying in local neighborhoods instead of central tourist districts.

Tools that are available to address these information needs all fall short. Traditional guidebooks from Fodor's, Frommer's, and Lonely Planet give people information about those central tourist districts and sights to see. Yelp and Foursquare give people information about the businesses, the bars and restaurants and shops, in an area, but travelers can't understand how the whole neighborhood feels just from those venue-specific details. Cities gather statistics - and indeed, are releasing open data more than ever before - but numbers also fail to convey a neighborhood's culture. Finally, occasionally travelers can learn local vernacular descriptions, but these are often shallow. For example, "Lawrenceville is the cool neighborhood" or "South Side is the party neighborhood."

At the same time, people take pictures differently than they used to. Most adults in the US have smartphones [5] and therefore instant access to a camera. Applications like Facebook, Flickr, and Instagram have turned photo sharing from a niche practice into a common one. Beyond photos, people have been sharing more on Twitter, Foursquare, and other social networks than ever before. Furthermore, people's smartphones usually have GPS, enabling users to quickly tie

their photos to real-world locations. As a result, there is a large quantity of public geotagged photos, as well as tweets, reviews, and other kinds of posts. One estimate has at least 500 million tweets sent per day [46], and a recent dataset release from Yahoo included 100 million public photos, of which about half are geotagged [89].

In this thesis, I argue that these photos and other posts, all tied to locations, could help travelers understand the cities they are traveling to. After all, pictures are rich sources of media that can give people lots of information: the proverbial "thousand words." Social media posts have other advantages over traditional travel guides as well. They are scalable, so it is feasible to build guides cheaply to cover any neighborhood in almost any city, and they are democratic: the resulting view of the neighborhood is controlled not by a publisher or elite critic, but by everyone adding their own experiences.

If it is the case that social media can be valuable for travelers, it sheds light on an even bigger question: what do geotagged social media posts tell us about the cities where they are produced? A wide swath of research has tried to infer everything from how socio-economically deprived an area is [71] to how attractive and magnetic it is [67], all from public social media.

In this document, I detail a research through design exploration into three questions:

- What do creative tourists mean by "getting a feel for the city"?
- How can social media help them achieve this goal?
- What can social media tell us about our cities and neighborhoods?

To do so, I first studied why people post geotagged social media data by analyzing public geotagged data and surveying 78 frequent geotagged tweeters and 400 geotaggers across six social media services. I then built a model of tourist information search based on interviews with 14 travelers and 490 survey responses. Based on this model, I designed a Neighborhood Guides web app and evaluated it with 10 participants to try to address their needs. I then tested six different sources of photos in a Mechanical Turk study to determine which photo set best represented their neighborhoods. Finally, with an improved set of photos, I conducted a user study with 21 participants on an improved version of the Neighborhood Guides web site, trying out different variants to best understand what they were seeing in these guides and how it could help them.

The contributions of this thesis are therefore not the construction of the Neighborhood Guides website, but the insights that its iterative design process has provided to these research questions above. I found that creative tourists want safety, convenience, liveliness, aesthetic appeal, and the ability to live like a local. Social media, particularly a diverse and well-organized set of photos selected using their annotations and contents, can help creative tourists find these dimensions they want. In this way, social media can offer one lens into the city: it shows the idealized city, not the "realistic" one, but this is what these tourists want.

These findings can help companies, or indeed local governments, design travel guides for future travelers. This will make it easier and more fun to travel to big cities, but also more fun to travel to mid-sized or small cities that currently do not get as much tourist attention. With international travel destinations like Paris and Venice losing character due to a deluge of tourists and smaller

but worthwhile cities like Cleveland and Atlanta needing ways to attract investment, this could benefit the entire tourism industry. Furthermore, this guide could be useful for planners beyond travelers, especially for neighborhoods that are growing or developing and want to be like other more popular neighborhoods.

Point-based guides and statistics can only go so far to help us understand crucial aspects of a city's culture, and travelers nowadays want to know that culture more and more. This tool will help people deepen their understanding of cities, and help researchers learn about cities as well. The rest of the chapters in this paper will be structured as follows:

2. A review of related work regarding geotagged social media and tourism
3. Three studies about why people geotag: One observational study of Twitter and Flickr users and two surveys (N=78 and N=400)
4. Two formative studies (14 interviews and a survey of 490 people) leading to the development of a model of creative tourist information search
5. Design of the Neighborhood Guides prototype and an introductory (N=10) user study
6. Mechanical Turk study of the most representative photo set for a neighborhood (N=200)
7. Final user study with a revised Neighborhood Guides prototype (N=21)
8. Discussion combining findings from all the studies, and directions for future work
9. Conclusion

In summary, I show three research contributions: first, the model of creative tourists' information needs; second, examples of ways that social media can help tourists; and third, evidence that geotagged social media shows an idealized view of cities.

2 PRIOR WORK

My work seeks to answer questions about three topics: what modern travelers really want, how social media can help them achieve their goals, and what social media says about our cities. In this section, I will review related work along these three dimensions, and then describe how this work will extend them.

2.1 CHANGES IN URBAN TOURISM

While urban tourism was not a focus of early tourism research, it has recently become a growing field [23]. Travel in previous decades had meant traveling to beaches, beautiful natural sites, or resort towns; but in recent years urban tourism is the fastest growing segment of the tourism market, growing by 47% between 2009 and 2014 [11]. The character of urban tourism is changing as well as the volume: new urban tourists want to "experience and feel a part of everyday life" [57]. Furthermore, they seek to have an active hand in co-creating the experiences, rather than passively paying for and absorbing an experience [6]. Lists of sights to see and experiences to buy no longer suffice.

2.1.1 AUTHENTICITY

When modern tourists travel to a city, they are often looking for an authentic experience of that city, rather than a manufactured diversion. The search for authenticity in tourism has a long history dating back at least to the 1970s [56]. This early work suggested that places had a "front stage" and a "back stage", like Goffman writes about in people [25]. There are different types of "front stage" and "back stage"; a shop or a bank might be the most "front stage", living with someone in their home might be very "back stage", and experiences like a tour of orchestra practice spaces or a cooking class in someone's home might be at an intermediate level. MacCannell suggests that all travelers want to get closer to the back stage, but this is clearly not strictly true; after all, many people travel explicitly to very "front stage" locations like Disneyland. Later writers have partially resolved this tension by talking not about how "authentic" a place is, but by how authentic an experience in a place is [95]. I will discuss the authenticity tension further in Chapter 4.

Regardless of the details of travelers' exact definitions of "authenticity", recent developments have aided their search in new ways, particularly with regard to lodging. Because hotels historically clustered in a few areas of cities, like downtown and near airports, they cannot show travelers all the sides of a city they may want to see, so travelers are turning to alternatives. The peer-to-peer lodging rental site Airbnb, for example, has become a popular, and more "authentic", way for travelers to rent rooms in residential parts of town [82, 96]. Similarly, Couchsurfing allows users to stay with locals for free (often on their spare couch, hence the name) [96]. As

urban tourists change from "mass tourists" to "cultural" and "creative" tourists [76], "mass" lodging no longer suffices either.

New urban tourists want to stay in interesting residential neighborhoods and spend time "wandering about", "taking in the city", and "getting among the people" [6]. To do this, they need guides to areas, not specific venues. Urban tourism, unlike other forms of tourism like "sun and sand" tourism, depends on the serendipity and spontaneity that results from getting to know neighborhoods, and on the individual's ability to co-create their experience. Current tools help people discover points, not overall pictures of parts of the cities.

2.1.2 ANTI-TOURIST IDENTITY

One interesting characteristic of modern tourists is to think of themselves as particularly unique and unlike "other tourists." For example, in one study of German tourists in Norway, 89.5% of respondents thought of themselves as "nontypical tourists" [69]. Many have investigated this kind of "anti-tourist attitude"; Jacobsen provides a thorough overview [36]. This attitude, present since at least the 1960s, seems to be a reaction to the democratization of travel and the homogeneity of mass tourism. Travelers see themselves as largely outside the tourist role for a number of reasons: to maintain social status when outside their normal lives, to explore a heroic ideal of exploration, and to assert identity in a group that they cannot quite afford. There is also a perception that tourism hurts a local area, as in Kreuzberg, Berlin, where travelers will go out of their way to convince themselves that they are not contributing to the gentrification and commercialization of the area [23].

Creative tourists are more likely than others to perform this anti-tourist identity. Edensor describes off-the-beaten-path travelers like backpackers preferring like to distinguish themselves from package tourists and from each other, adopting various non-conformist tourist performances to do so [22]. As a result, to visit a place that is "touristic" would be not only frustrating but even damaging to visitors' performance of their chosen identity. Creative tourists would reflexively recoil from something that they thought would put them "on the beaten path", so it is important to focus any efforts to design for them with this important identity characteristic in mind.

2.2 PROVIDING RECOMMENDATIONS TO TOURISTS

Using social media to help tourists is not a new idea. Since the early 21st century, researchers have tried to use the abundance of social media data to recommend things for tourists to do. Work in this vein includes recommendations of restaurants [34], shops [84], travel routes [49, 65], attractions and points of interest [24], and destinations [29]. These all use social media and user-generated content such as user locations, so continuing in this vein seems like a logical choice. In addition, sites like Yelp and Foursquare have dozens of user reviews, so aggregating reviews and recommending the most highly-rated spots seems like a natural solution.

However, this approach has three shortcomings. First, people need to know why they are recommended each place. It would be rare for tourists to set out on a trip solely because an algorithm recommended it. Second, they solve problems that are already solved by Yelp and Foursquare: finding a restaurant or a point of interest by consulting one of these guides is easy. Finally, these works neglect the changes in urban tourism discussed recently. A recommendation algorithm

will likely push more people to the top destinations, which then become overcrowded and no longer as enjoyable. Instead, we need guides to let people explore places on their own time and create their own connections to them. For these reasons, I decided not to continue in the vein of providing discrete recommendations like these researchers have done.

2.3 What Social Media Says About Cities

Because of the shortcomings in the straightforward recommendation approach to computational tourism, and the need to build browsable guides instead of pinpoint recommendations, it is valuable to take a wider view. Plenty of researchers have attempted to deepen our knowledge of cities with social media in a number of ways. The rest of this section will survey those approaches.

2.3.1 Delineating Neighborhoods and Regions of Cities

One useful application is in finding the boundaries of regions. Most large US cities have official neighborhood boundaries, but these have several problems. They may be out of date, there may be multiple conflicting definitions[1], and they often fail to reflect the reality of human behavior. These are often politically or financially motivated, like "NoPa" and "Lower Nob Hill," which San Francisco realtors use to convince affluent clients that they are not in the less desirable Western Addition or Tenderloin neighborhoods.

Instead of settling for these confusing and misleading neighborhood divisions, recent work has been able to find reasonable neighborhood boundaries based on human behavior such as Foursquare checkins [18, 97] or tweets [94]. This can reveal aspects of neighborhood life that is otherwise hidden, such as a neighborhood that contains two mostly-separate social sub-neighborhoods. This idea has also been extended beyond neighborhood bounds to delineating less formal locally characterizing regions (like "red light district") based on photos [88].

My work differs from these in terms of goals. While I see the need for a better definition of neighborhoods, I am not trying to do so myself, instead starting with neighborhoods as defined by local government. I have chosen to do this because official neighborhoods instantly recognizable (people are more likely to have associations with "Squirrel Hill" than with "Neighborhood 34") and in order to scope the project reasonably. Also, while there are border cases where people have strong opinions about the exact definitions of their neighborhoods, these are likely rare and will not affect results much.

2.3.2 Describing Regions by Summarizing Social Media Posts

Beyond finding the boundaries of regions, researchers have discovered ways to understand those regions based on social media. Because the quantity of social media is huge, these approaches can be thought of as ways to summarize or model all of these posts.

Photo-sharing sites, particularly Flickr, have been well studied, due to the volume and richness of their posts. Some of this research has been driven by practical concerns, like the need to show

[1] for example, San Francisco's "SF OpenData" Portal offers the following:
https://data.sfgov.org/Geographic-Locations-and-Boundaries/Analysis-Neighborhoods/p5b7-5n3h,
https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find-Neighborhoods/pty2-tcw4, and
https://data.sfgov.org/Geographic-Locations-and-Boundaries/Realtor-Neighborhoods/5gzd-g9ns

photos on a map. Toyama et al [91] developed techniques including thumbnails, point markers, and isopleths to show how many photos existed on a map at a place before settling on a binning approach they call "media dots." However, these displays only show the number of photos, not their contents, so a series of other projects worked on summarizing photo content as well as density.

Some of this research works on finding a subset of photos that is representative of a larger set. Jaffe et al [37] addressed the problem of summarizing photo content by finding a subset of photos that would accurately summarize a larger photo set. They did this by clustering all of the photos and then ranking the clusters based on five criteria: tag distinguishability, photographer-distinguishability, density, image qualities, and arbitrary relevance factor (such as a search query). Kennedy et al [43] further developed the ability to find the "most representative" image from a set of photos using computer vision features such as SIFT. Crandall et al [15] did the same: finding the top N "interesting" places in each city and a "canonical" photo from each.

Besides investigating photo contents, researchers have investigated ways to summarize the textual tags that users add to their photos. Ahern et al [4] and Jaffe et al [37] describe the World Explorer/Tag Maps project, which summarized a series of photo tags into "representative tags" for a region. Kennedy and Rattenbury expanded this to describe semantics of places and events [43, 72], while Kafsi et al further expanded it to understand which tags are locally relevant, which are city-level, and which are country-level [42].

Summarizing textual content, like tweets, is somewhat easier because there is less total information, so one can use a simple method like a word cloud (at least as a supplementary tool) to get a sense of a large corpus of words [59]. More intelligent methods have been used for tweets, for tasks like event detection [48] and location modeling [44]. Importantly for neighborhoods, though, Hao et al approach high-level neighborhood modeling in another interesting manner, creating Location-Topic Models based on what users write in travelogues [29].

Finally, neighborhood comparison [51] offers a way for people to understand neighborhoods in a new city based on neighborhoods that they already know. This can help people talk about imprecise or unnamed characteristics of neighborhoods: they may not know what they like about their home neighborhood, but they know that they want to find someplace like it.

This work is quite interesting and I hope to reuse and extend it. I will do so by repurposing their results; where the researchers in [37] delivered an algorithm to pick a subset of photos as a final product, for me this is only one input into the application that I'm building. Similarly, neighborhood comparison was the end result in [51], while it is only one feature in the application that I am building. In summary, my work will be more about human needs in tourism, rather than about deep algorithmic details.

2.3.3 DESCRIBING CITY RESIDENTS

Another set of research focuses on describing people who live in a certain area instead of the area itself. Some of this work involves general studies of mobility, routines, and urban dynamics [7, 45, 61], indicating where people go when and how far they travel. For example, Komninos et al [45] used a network of listening stations around Patras, Greece in order to see Foursquare

checkins around the center of the city. They were then able to quantify business traffic throughout the day, and report where the "hot" areas of town were, so that locals and tourists could visit or avoid them. Naaman et al used a similar approach on Tweets to show the standard daily patterns in cities, and then show whether certain keywords followed similar patterns, to reflect mood and happiness. Besides being useful in itself for business owners and transportation planners, this mobility research can in turn describe well-being of a region [50]. Some also skip the middle step and directly predict socio-economic well-being from sentiment analysis of social media posts [70]. These studies can be useful, albeit noisy, indicators over large urban areas.

Some researchers instead focus on demographics of the users. This work can be low-level, such as showing the gender makeup of commuters [55], or higher-level, indicating "topics" of users based on their interests [40]. While these are both valid approaches, I chose not to focus on them because they do not serve the same goal of helping travelers. Demographics and mobility are both too vague for travelers to make use of in a meaningful way.

2.3.4 CHALLENGES IN THESE APPROACHES

While these approaches all have valid uses and results, they contain some shortcomings. Goodspeed cites three main issues [27]:

- Content Poverty: social media contains broad data from many users, but it usually fails to contain deep enough data to answer any one particular question. For example, a transit researcher would lose the richness of traditional travel surveys, instead only knowing "this person was near this point at this time."
- Espoused Theory vs. Theory-in-use: social media data either tells what people are saying or what they are doing, but not both, which hampers our ability to understand their reasons for their actions. Especially in a public forum, people may not be doing exactly what they say they are doing. Their locations may even be spoofed, as many platforms allow. When we are using public data, we must realize that the data is only what people want to present of themselves.
- Positivist Assumptions: studying social media often implicitly assumes that the sample of society that is shown on these public posts is representative enough of the world to answer the researchers' questions. For example, if a researcher wanted to show where people go at night in order to prioritize late night bus routes, they would have to contend with the fact that the data only reflects the subset of users who tweet.

In short, social media data is broad but not deep; we should not conflate the size of data available (often in the terabytes) with the usefulness of that data. As one example, Hecht and Stephens [32] show that this data is biased towards urban areas and underrepresents what happens in rural places. As a result, we cannot simply use existing work directly to show tourists exactly what the city is like.

In summary, this chapter presents three types of work. First, tourism research shows that the way people travel is changing and that new tools, guides, and techniques are necessary in order to support them. Second, existing recommendation systems offer one avenue to help these tourists,

but this is unsatisfactory. As a result, we turn to the third area of research: how social media helps us understand cities now, and how these techniques may be applicable in the future.

3 WHY PEOPLE GEOTAG

Before analyzing social media posts, it is important to know about how these posts are created. This is a necessary first step to analyzing public social media data, because it offers context about what the data means. For example, imagine an Instagram picture of a meal. If it was created by someone while they were posting a Yelp review, it would signify something about the meal: perhaps that it was worth reviewing. If it was created by someone who automatically uploaded all their photos, and they were currently trying to track what they ate, it would mean much less. Therefore, Alex Sciuto, Zichen Liu, and I set out to study why people post public geotagged social media posts, in order to inform the rest of the thesis work.

In this chapter, we conducted three studies. First, we investigated public geotagged tweets, finding that geotagging is toggled more than we expected, that Twitter changed its interface in ways that affected our data, and that job-posting spam bots became a nuisance. Second, we surveyed 78 of the most frequent tweeters in our data set to understand why people choose to add their locations. Third, we conducted a larger survey of 400 people across multiple platforms to see their motivations for their last geotagged post. We found that people do usually geotag consciously, at places far from home, and at places that they have not been to very often.

This research helped answer my second and third research questions: "Can social media help travelers find what they want?" and "What does social media tell us about our cities and neighborhoods?" The work in this chapter was published as *State of the Geotags: Motivations and Recent Changes* at the AAAI International Conference of Web and Social Media (ICWSM) in 2017 [86].

3.1 BACKGROUND AND MOTIVATION

Past work has investigated people's behaviors on location-based social networks (LBSNs) like Foursquare, Facebook Places, Dodgeball, and Gowalla, to learn who shares their location and why [28, 53, 85].

However, many non-location-based social media services–including Facebook, Instagram, Twitter, Snapchat, and Flickr–also allow users to add their location to a photo, video, or text post. One study estimated 600,000 geotagged posts per day from a 10% sample of Twitter [52]. Flickr, meanwhile, has released a publicly available dataset of 49 million geotagged photos [90]. A wide field of research has sprung from this wealth of publicly available geotagged posts, such as understanding demographics and social dynamics in cities [60], finding home locations of individuals [41], and inferring likely friends [16].

For LBSNs, there is often a clear reason as to why someone checks-in to a location or tags something. However, there is currently little understanding of what people geotag on these popular social media sites not centered around location, and why. Understanding what is being geotagged

and why can have implications for research. For example, finding people's home locations based on their check-ins could be very easy if people geotag mostly at home, or very difficult if they only geotag when traveling. As another example, models about people's mobility patterns and social dynamics will be very different if they are based on commute data or weekend shopping and errands data.

It is easy to assume that geotagging in social media is similar to that for LBSNs, but without explicit investigation it can be difficult to know for cetain. To address this problem, we conducted a series of studies to understand whether people use geotagging in social media similarly to the way they use location-based social networks. We analyzed 4 million public tweets and 49 million geotagged Flickr photos, surveyed 78 frequent Twitter users, and followed up by surveying 400 geotaggers across six social media services. We found that most earlier findings in Foursquare ring true in other social media: people geotag consciously and intentionally, they geotag in uncommon places, they primarily do so to communicate and show where they've been, and they geotag soon after being at the place.

However, our analyses uncovered several new findings. We found that most Twitter users geotag consciously and turn geotagging on and off frequently, but many Twitter users were inadvertently geotagging, or geotagging more precisely than they thought. We uncovered a UI change that addressed this issue, while also causing people to add coordinates less frequently and add place names more frequently. We also discovered that the coordinate geotags that remain tend to have more hiring-related spam.

Our findings have several research implications, given how often researchers use geotagged data. It is important that the research community not misunderstand what people are providing when they publish geotagged social media posts, and it is important that we minimize the impact of spam and other quality problems. Knowing why people geotag also helps application developers better customize their software.

To support these research and development implications, this paper offers two contributions. First, we show confirmation and elaboration of earlier findings, and generalization from Foursquare to other social media. Second, we expose a number of changes that have occurred as location sharing has matured.

3.2 DEFINITIONS

In this chapter, I use the term "geotag" to mean "a location added to a social media post." We use "coordinate geotag" to mean "an exact latitude-longitude coordinate added to a post." A related concept is a "placetag", or a tag referring to a plain-text location. For example, a post at the Eiffel Tower could contain the *placetag* "Eiffel Tower," "Paris," or even "France." It could also or instead include the *coordinate geotag* (48.858, 2.295).

We refer to "checking in", as in Foursquare and other location-based social networks, as a separate but related act. In Foursquare, one opens the app primarily to share one's location. When geotagging a tweet, photo, or other post in social media like Twitter, Flickr, Instagram, and Facebook, however, the content of the post is usually the primary motivation, while location sharing is usually secondary.

3.3 STUDY 1: ANALYZING PUBLIC GEOTAGGED DATA

To understand what and why people geotag, we chose to start by analyzing publicly visible data from Twitter and Flickr, primarily because they offer the largest public data sets of geotagged posts.

We started by collecting geotagged tweets via Twitter's public streaming API. We chose to start in Pittsburgh because it has a wide variety of users and because of our team's high familiarity with the area. We selected all coordinate-geotagged tweets within 0.2 degrees latitude and longitude from the center of Pittsburgh, forming a 34km x 44km rectangle with corners at (40.241667, -80.2) and (40.641667, -79.8). We omitted tweets that listed an area (like "Pittsburgh") but did not contain a latitude-longitude point. We began gathering data in January 2014, and by May 2016 we had about 4 million tweets. We also gathered data in 12 other cities, mostly around the United States[1], to verify any results we found on the Pittsburgh data set. These other cities' data covered a shorter time span (11-23 months) but still the same order of magnitude of tweets, from about 1 million in Austin to 11 million in London, totaling 60 million tweets.

We also examined the YFCC100M dataset [90] to gather information about geotagging on the photo-sharing site Flickr. This data set contains metadata for 100 million photos and videos that are shared publicly with a Creative Commons license. Of these 100 million photos and videos, about 49 million are geotagged.

3.4 STUDY 1 RESULTS

3.4.1 PEOPLE OFTEN TOGGLE GEOTAGGING

In Twitter's mobile app, users can choose to geotag or not, but the default is whatever was set last. If a user geotags one post, the next one will be geotagged as well unless the user turns it off. As a result, we had initially assumed that geotagging was a setting people would mostly leave on or off; that they would decide to geotag or not to geotag and then apply that to all of their social media. However, this was not the case.

We selected a random sample of 3406 users from our data set and collected all of their public tweets, geotagged and non-geotagged. We sampled users because Twitter's API has rate limits of 180 requests per 15 minutes, and because it only supports collecting up to 200 tweets per request. As such, collecting all tweets from all 68088 users would have taken prohibitively long.

For each of those users, we sorted their tweets in chronological order, then counted a "toggle" every time they had a geotagged tweet followed by a non-geotagged tweet, as this likely indicated they had made a conscious choice to geotag or not geotag something. We only used tweets from their most frequent tweet source (such as "Twitter for iPhone" or "Twitter Web Client") to avoid counting false "toggles" caused by them, for example, tweeting from their phone then tweeting from their computer. We found that most people in this sample toggled geotagging relatively regularly, and only a minority (40.1%) toggled less than 1% of the time, as we would expect if they were geotagging automatically. Figure 3.1 shows the distribution of how many of each

[1] Austin, Chicago, Cleveland, Dallas, Detroit, Houston, London, Miami, Minneapolis, New York, San Francisco, and Seattle

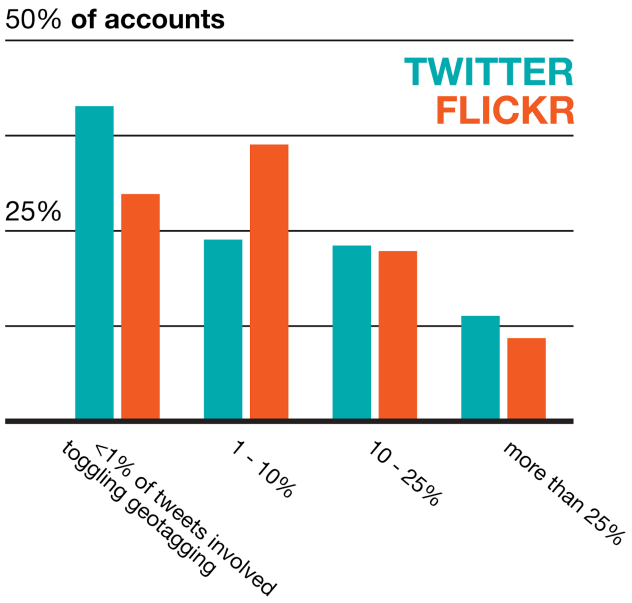


Figure 3.1: Percent of tweets and photos in which users turned geotagging on or off. Tweets are taken from a random sample of 3406 users out of all 68088 users who had geotagged at least once in Pittsburgh; Flickr photos are taken from the YFCC100M dataset. The "Less than 1%" group can be considered those who always or never geotag; it is a minority, as most people toggle geotagging at least occasionally.

users' tweets were preceded by a toggle, while Figure 3.2 shows toggling patterns for three example users.

We saw similar patterns in the YFCC100M dataset. Out of the 581099 Flickr users in the data set, 214598 (36.9%) had posted at least one geotagged photo. For these users who have geotagged at least once, we counted toggles the same way as for Twitter users; results are shown in Figure 3.1. Seeing roughly the same pattern as in Twitter gives us confidence that geotagging on social media platforms is a conscious choice, not automatic.

3.4.2 Changes in geotags over time

Although it was not a primary research question, an interesting finding emerged about the pattern of geotags over time. In the YFCC100M dataset (see Figure 3.4), the overall count of photos is rising in frequency, though geotagged photos are tapering off. However, [38] found the percent of photos with geotags in Flickr is increasing. One possible explanation is that the decline in geotagged photos is an artifact of the process of creating the YFCC100M dataset (or people's willingness to license photos as Creative Commons).

In our Twitter data, however, we noticed a sharp dropoff in the number of geotagged tweets available after about May 2015. This was consistent across all of our cities; Figure 3.3 shows the dropoff in a few example cities.

This is due to a UI change Twitter made in April 2015, when announcing a new partnership with

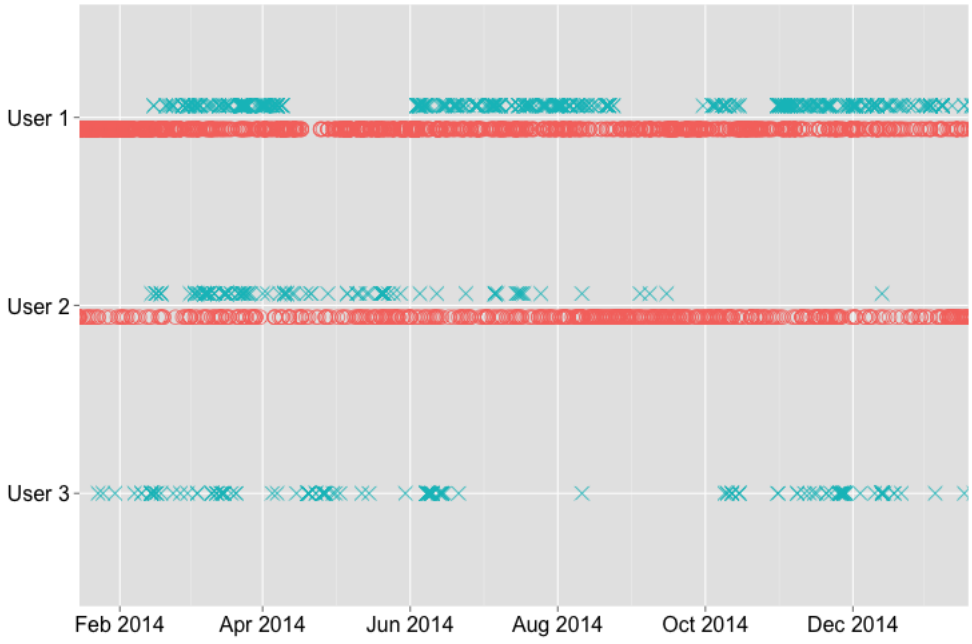


Figure 3.2: All tweets from three representative users. Blue Xs represent geotagged tweets while orange Os are non-geotagged. User 1 geotags while traveling, User 2 tags because of a Wordpress plugin, and User 3 sees no reason not to geotag.

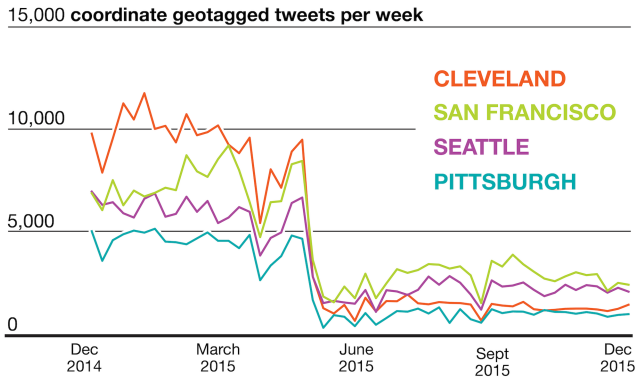


Figure 3.3: Counts of coordinate geotagged tweets from different cities. The sharp dropoff around May 2015 is due to a Twitter UI change: placetagging, not coordinate geotagging, is now the default.

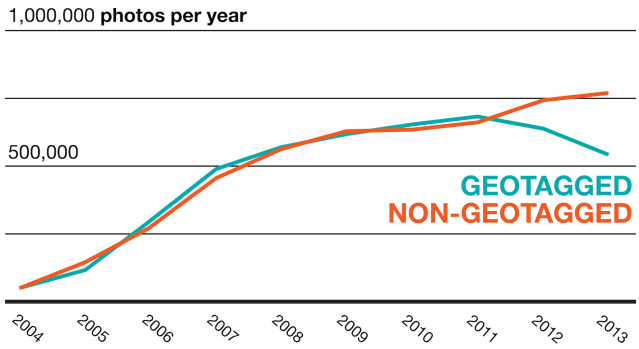


Figure 3.4: Counts of geotagged and non-geotagged photos from the YFCC100M dataset. Note that, while non-geotagged photos are growing steadily, geotagged photos may be falling.

Foursquare[2]; placetagging, not coordinate-geotagging, is now the default. As we will discuss in Study 2, while this reduced the data available to researchers, it also removed a source of confusion and accidental privacy leaks.

3.4.3 How many distinct places do people geotag at?

We calculated how widely people's geotag distribution varied. After excluding all accounts with fewer than 20 geotagged tweets, we rounded each geotag to the closest 0.001 degree latitude and longitude (creating bins about the size of 1-2 city blocks), then counted how many places each account had posted a geotagged tweet. We found that most people had between 1 and 35 places (median = 18, 3rd quartile = 35, 95th percentile 103.8; see figure 3.5), but of course this depends on the number of tweets. We also calculated the number of tweets per place, finding that people usually tweet about 4 times in a place, though this varies widely (1st quartile = 2.3, median = 3.9, 3rd quartile = 7.0, 99th percentile = 102.6).

[2]https://twittercommunity.com/t/foursquare-location-data-in-the-api/36065

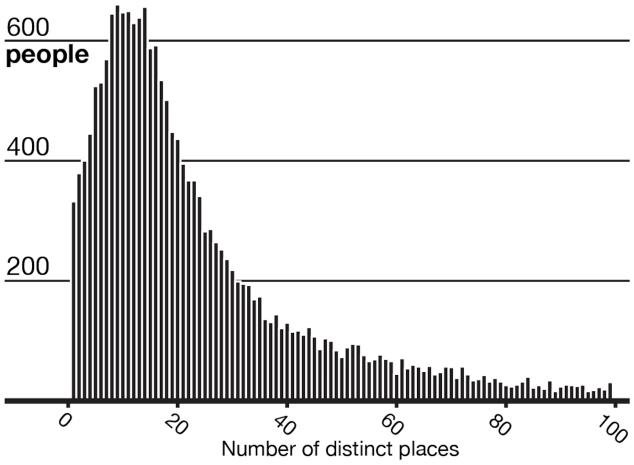


Figure 3.5: Number of distinct places where each Twitter user in our Pittsburgh data set (with at least 20 tweets) geotags. Notice how many users tweet only in one place.

City	One-place accounts	Job bots	%	Multi-place accounts	Job bots	%
Pittsburgh	331	73	22.1	19,214	179	0.9
SF	394	103	26.1	31,777	217	0.7
Seattle	259	78	30.1	15,571	203	1.3

Table 3.1: Distribution of job posting bots on Twitter in Pittsburgh, San Francisco (SF), and Seattle. "One-place accounts" are accounts that post in only one location (rounded to the nearest 0.001 degree latitude and longitude). "Many-place accounts" post in multiple locations. Not all job posting bots post in one place, but a large percentage of one-place accounts are job bots.

One surprise we found is that some accounts geotag repeatedly in the same place. We inspected a random sample of 50 of the one-place accounts in the Pittsburgh area and found that about 15 accounts were bots that only posted job listings, 7 accounts had been deleted or protected, 1 was a bot that tweeted weather reports, and 1 was a bot that tweeted NHL hockey scores. While we could not detect every bot, all of the job posting bots included "job", "career", "work", "join", or "tmj" in their name (and none of the other accounts did), so we could easily scan for other bots in the full data set. We found that a large percentage of one-place accounts were job posting bots; complete statistics are shown in table 3.1.

Of course, there are plenty of spam accounts on Twitter besides job posting bots. Some of them, like realistic-looking accounts made to promote a product, are difficult to filter out. However, we point out the job bots to show one easy way that researchers who are analyzing geotagged tweets can easily remove a large quantity of tweets that may not be relevant to their purposes.

1	What is your Twitter username?
2	Did you know that you've posted geotagged tweets in 2014? (answers: yes, no)
3	Describe the first time you geotagged a tweet. What caused you to decide to add your location?
4	Do you still geotag your tweets? (answers: Yes, always; Yes, sometimes; No; I'm not sure)
5	If you currently geotag your tweets sometimes, describe a recent tweet that you decided to geotag.
6	If you currently geotag your tweets sometimes, describe a recent tweet that you specifically decided NOT to geotag.
7	Are you worried about privacy implications of geotagging your tweets? (Answers: yes, very worried; yes, slightly worried; no, not very worried; no, not worried at all)
8	Why or why not?
9	Which Twitter client do you use most often?
10	Did this survey cause you to change your choices about geotagging?

Table 3.2: Questions in Study 2. All responses without answer choices given were free-response.

3.5 STUDY 2: SURVEY OF TWITTER GEOTAGGERS

In Study 1, we found that social media users often toggled geotagging, which suggests that people may have nuanced views of privacy and sharing. However, there have also been news articles indicating that people sometimes accidentally shared geotagged media too. We were interested in probing these behaviors more. Towards this end, we conducted a survey of Twitter users.

In November 2014, we compiled a list of our users, sorted by the number of times they tweeted in our data set. After our study was approved by our IRB, we recruited 4119 participants to take a survey by tweeting a link to them. We started from the most prolific tweeters in order to make sure we had active users. A total of 78 responded and were paid with a $5 Amazon.com gift card for participating. (While we wish we could have had a higher response rate, we considered it appropriate for an exploratory survey.) Survey questions are listed in 3.2. Free-response survey questions were analyzed using affinity diagramming, as described in [8]. This technique, in which all main points from responses are printed out on post-it notes and grouped iteratively according to main themes, allowed us to find higher-level themes that emerged from the data in a bottom-up manner.

We intended this as a preliminary study; because we had just been studying users' public tweets, hearing from them directly would be helpful. However, we also realized that only studying on one platform limited our results. While we started from the most prolific users and recruited down the list, this reflects a diverse array of active users: our users had between 57 and 2766 tweets over the course of our one-year time period (median=293).

3.6 Study 2 Results

3.6.1 Why they geotag

The most popular reasons people gave for geotagging their tweets were to communicate with and to show off their travels and events to followers. Of our 78 participants, 17 people described geotagging their tweets at an event, 9 described geotagging while traveling, and 18 described a more general desire "To show my followers where I am." This latter set of users described choosing whether to geotag each tweet, rather than simply leaving it on. This diversity of reasons inspired us to look deeper into reasons behind geotagging, which we do in Study 3.

3.6.2 Some users did not know they were geotagging

Surprisingly, nine participants reported being unaware that they were posting geotagged tweets, while six more reported accidentally turning it on at some point and then consciously deciding to leave it on. Four were persuaded to start geotagging by an app and 10 decided to start geotagging on a whim or out of curiosity.

One major reason that people may be unaware of their geotagging is the presence of third-party apps that post geotagged tweets. Two users mentioned that they cross-post geotagged Instagram photos to Twitter, while a third uses a Wordpress plugin that cross-posts blog posts. This user was surprised to learn that she had been geotagging at all. Third-party Twitter clients are also possible causes: one person who used the Tweetbot client (an app for reading and posting tweets) mentioned that Tweetbot enables geotagging by default.

3.6.3 Some did not know their geotags' precision

Most participants expressed some concerns about privacy, including vague feelings that they did not want to tell the world where they were (15 participants), or that they specifically did not want the world to know when they were not at home (8 participants). Several explicitly mentioned potential burglaries. These concerns echo previous location privacy findings [92].

However, 20 participants expressed very little concern about privacy. Worryingly, 12 of these participants expressed belief they were only sharing broad city-level locations, and thought that nobody knew their exact location. However, everyone in our data set, including these 12, had posted public tweets with precise coordinate geotags.

Upon further investigation, we found that the Twitter mobile app showed a confusing user interface: it appeared that users would be posting high-level tags (like "Pittsburgh, PA") when instead the actual latitude-longitude point was stored with the tweet. As noted earlier, the Twitter mobile app's user interface has since been changed to use placetags.

3.7 Study 3: Cross-platform geotagger survey

While the survey on Twitter users raised some new interesting questions, it did not fully answer the question of why users geotag. In addition, it focused only on Twitter users. We wanted to increase our sample, as well as broaden it to include other social media users.

Our primary research question for this survey was "Why do people geotag?" Having read many

Motivation	Theoretical basis
To show that I was at a cool, amazing, special, or popular place	Common in [28, 54]; handles "Impression management" theme without introducing jargon into the survey
To keep track of this place for later on	One of the 5 factors in [53]
To promote this place to my social network	Related to the "Place discovery" factor in [53]; related to self-presentation discussed in [14, 28]
To coordinate with my friends for activities	Part of the "Social connection" factor in [53]
To keep friends/family updated on what I'm up to and where I'm at	Part of the "Social connection" factor in [53]
It's automatic/I always have geotag on	This does not appear in LBSN research because it doesn't apply to LBSNs, where checking-in is the app's purpose.
Other: _____ (free response)	

Table 3.3: Answer choices for "What are your motivations for geotagging?" in our Study 3. Participants could check all options that applied to them. (Only the "Motivation" column was shown to them.)

papers about why this topic in location-based social networks [14, 28, 53, 54], we wanted to see if their findings about why people check in can be replicated in geotags in non-location-based social networks. As such, we asked them to pull up their most recent geotagged post in any social media and asked them, "Please explain in detail why you geotagged," with a free response answer. We also asked, more generally, "What are your motivations for geotagging? (check all that apply)" and offered the choices in table 3.3.

We recruited 406 people from Amazon Mechanical Turk to ask them about their geotagging practices and their most recent geotag on their most used social media service. We recruited participants who were in the United States and had previously geotagged at least once. We asked about the content of their most recent post with geotag: what type of post it was (text, picture, etc), what it was about, why they posted it, and how meaningful it was. We asked about the geotag: what the geotag was, how precise it was, how they would describe the place they geotagged, why they geotagged, what their motivations for geotagging are, how often they go to the place, how far it is from where they live, and how long they waited to post it. The study took about 5 minutes and participants were paid $1. Six people's responses had to be removed as their free responses indicated they were not paying attention, leaving us with 400 valid responses.

Time passed	Percent of respondents
Happened at the same time	40.75%
Within an hour	28.75%
Within a day	20.5%
Within a week	6.75%
Within a month	2.0%
More than a month	1.25%

Table 3.4: Responses to "How long did you wait between having the experience and posting?" in Study 3.

3.8 STUDY 3 RESULTS

For reasons of space and conciseness, we will not report on all of the survey questions, instead highlighting some of the most surprising and relevant findings here.

3.8.1 WHEN DO THEY GEOTAG?

Most of our participants reported geotagging in the moment, mostly within an hour of the photo or event they are sharing. If not, they usually geotag by the end of the day: 69.0% geotag within an hour, 89.3% within a day (see table 3.4 for details). However, a substantial amount waited, for reasons like wanting to 'settle down (at hotel)' , 'find phone service signal', or posting upon friends' request.

This suggests a slight difference between geotags and check-ins, due to the concept of "check-in transience" introduced by Guha and Birnholtz [28]. Check-ins have a short lifespan; it would rarely make sense to check in later in the day or week. However, about 1/3 of our respondents waited over an hour to share their geotagged post.

3.8.2 WHERE DO THEY GEOTAG?

Our participants mostly geotagged in places far from home, as shown in Figure 3.6. When we asked about the most recent geotag, only 11.9% of them were in the users' home or neighborhood, and 46.7% were in their home city. Johnson et al [39] found similar findings: that in Twitter, Flickr, and Swarm, only about 75% of posts are from "local" users, but ours are more extreme; depending on how one defines "local," anywhere between 11.9% to 46.7% of geotags are from "local" users. Furthermore, 70.0% of these geotags are from rare places: places they go every year, a few times so far, or this is the first time. This confirms the finding in [53] that many people are reluctant to check in at routine places, and the finding in [14] that users avoid checking-in to home and work because it can be annoying.

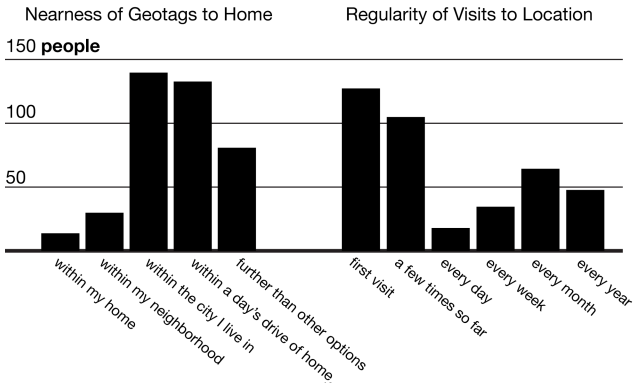


Figure 3.6: Where participants geotag. Note that most people geotag far from their home: only 11.9% tag in their home neighborhood, and less than half are in their home city. Most people also geotag in places that they visit rarely: 70% of these geotags happen in places that people go annually or less frequently.

3.8.3 WHY DO THEY GEOTAG?

We asked users twice why they geotag: once in general (responses are shown in Figure 3.7) and once about specifically why they geotagged their last post. The options shown in Figure 3.7 are based on previous research, as explained in Table 3.3. We can see that few people chose "other", which suggests that these choices explained people's preferences well. In addition, few people geotag automatically, which corroborates our finding in Study 1. The sharp difference in magnitude between the top two reasons and the rest adds some nuance to our knowledge: the most commonly cited reasons for geotagging are the social-driven ones ("show I was at a cool place" and "keep family/friends updated"), more than the purpose-driven ones.

We analyzed the free response question about why they geotagged their last post using affinity diagramming (as described in [8], as we did in study 2.) The categories that emerged almost all fit into one of the six choices shown in Figure 3.7. The two categories that didn't fit were "No reason" and "Application-driven"; we give some examples in Table 3.5. We note that these, too, have some precedent in the literature; users describe "checking-in" as something to do when bored and are motivated by game elements included in LBSNs [53].

3.9 DISCUSSION

3.9.1 GEOTAGS ARE POSTCARDS, NOT TICKET STUBS

Our studies brought us some findings that, unsurprisingly, agree with research on LBSNs. People geotag to show where they've been; keep their family and friends updated on their travels; record a place for later; or help family, friends, and strangers to find a place. People geotag at rare places more than routine places.

However, our studies add some depth and nuance to our understanding of geotagging. Showing off where they were and keeping family and friends updated rank as the most important reasons to geotag. In addition, we document for the first time that most people geotag consciously; they do

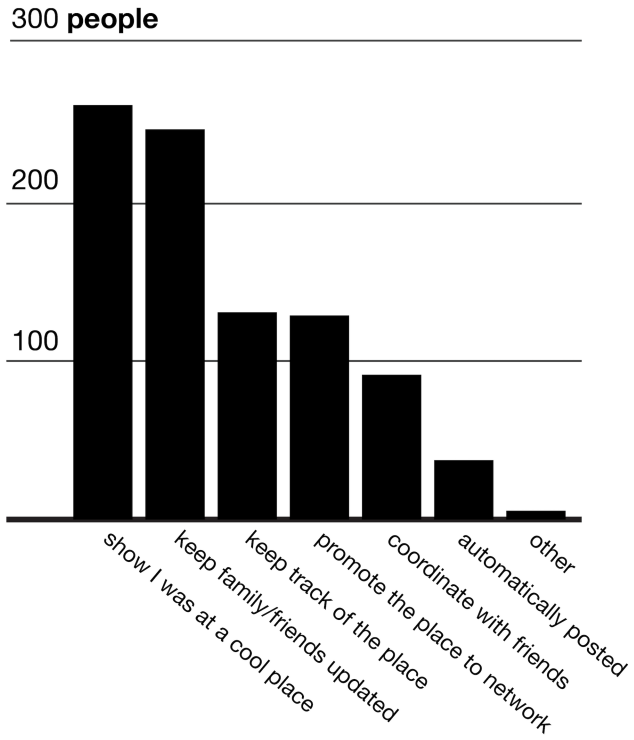


Figure 3.7: Stated motivations for geotagging. Participants (n=400) could choose multiple options.

No Reason
“No specific reason, just because I wanted to share the dealership where I was buying the car.”
“No particular reason other than I tend to enjoy seeing the geotags of my friends so I decided to include it myself.”
“Just because I wanted to, no reason in particular, just something to do.”
“I didn’t really think of it, I just kind of did it...doesn’t that explain most of what is on social media? Just because...”
Application-driven
“the geotag put the houston dynamo colors and logo over the picture”
“to get a discount off my melon boba tea.”
“To remind myself where I found the pokemon, as well as letting my friends know that it could be a good spot for the future.”
“...it’s how I find interesting people to follow and I’m sure its how I have gained followers too [on Instagram]”

Table 3.5: Participants’ given motivations for geotagging that fell outside the ones in Table 3.3

not simply set their phone to automatically geotag everything, as we saw in Studies 1 and 3. This was not an issue in LBSN research, because using an app like Foursquare without geotagging would not even make sense. But in non-location based social networks, it is important to know whether a geotag is a side effect of another action, like a ticket stub, or a consciously chosen artifact, like a postcard.

Our research shows that a geotag can be seen as a postcard: it shows that a person is at a certain place, it is usually used for social communication, and it is hard (though not impossible) to fake.

In addition, our studies uncovered three new important points:

- The landscape of geotagging is changing, from coordinate geotags to placetags
- Some people may not know that they're geotagging, or how precisely they're geotagging
- New types of spammers are becoming prevalent in public geotagged data

In the rest of this section, we will discuss the implications of these findings for different groups.

3.9.2 Implications for applications and app developers

There are clearly pitfalls to avoid when designing an application that involves users' locations. Many papers have documented the risks, e.g. [92]. In this paper, we documented two more risks: the possibility that users are coordinate geotagging when they think they are placetagging (as in Twitter before April 2015), and third-party apps that add geotags with people's knowledge. In this section, we propose ways to avoid these harms and improve users' geotagging experience.

Minimize automatic tags and coordinate geotags

To avoid accidental privacy leaks, social media software needs to be more careful when automatically geotagging participants' posts. For some apps, such as the old Foursquare check-in app (now Swarm), geotagging is the main purpose of the app, and so it makes sense to have automatic geotags. However, in Study 2, we also reported on two examples where people's mental models and expectations about geotagging did not match reality.

This point is obvious and straightforward. Less obviously, social media software might consider whether they want geotagging to be a sticky setting at all. That is, if a person geotags one post, should the next post be geotagged by default? Our results suggest no: people largely prefer to make a conscious choice about whether to geotag each post or not. This choice adds only slight overhead, and prevents potentially disastrous privacy leaks.

Another option to reduce privacy risks is to use placetags instead of coordinate geotags. All of our participants' main use cases could usually be handled by placetags, and coordinate geotags are often too precise, revealing more of a user's location than they want. Only the "coordinate with friends" and "keep track of this place for later" cases might require a coordinate geotag, and then only if it's in a wilderness area or other place without well-defined places.

Using placetags could improve the user experience in other ways as well, as coordinate geotags are usually hard for people to understand. Many services, like Twitter and Facebook, are already doing this well: users can select which granularity to placetag, whether it's the building or city

that they are in. They can also add their exact location if they want. A minor challenge is that coordinate tags can be generated solely on a smartphone through GPS, whereas placetags require network services and a large database that needs to be kept up to date to do lookups of place names.

HELP RESEARCHERS UNDERSTAND PLACETAGS

One notable downside of placetags, however, is their interpretability. If a researcher sees a placetag that says "Starbucks" without any finer grained information, how can they know which Starbucks location the user is at? Also, sometimes placetags represent coarse places, like "Singapore," but researchers interpret them as being finer-grained points. They sometimes transform a bounding box into one point at the center, which can have annoying or even disastrous consequences, as Shamma documents [79]. Services need to return geotags at different granularities: cities, polygons, or points. They also need to document why they are returning the granularity that they are: because the user chose it, because the user's GPS could not get an accurate reading, because of the user's default privacy settings, or whatever other reason.

While it would be computationally expensive to send an entire polygon with every social media point, it would be feasible to publish a gazetteer of places along with a social media API. Services may be tempted to send bounding boxes instead, but this could lead to other problems, such as Mapzen accidentally declaring Copenhagen part of Sweden [80].

Importantly, this is not an appeal to altruism; researchers are internal as well as external. Improving the comprehensibility of placetags for researchers will help a company's own analysts as well as the academic community.

3.9.3 IMPLICATIONS FOR RESEARCHERS

Many studies have treated geotag data as sensor data, without much regard for how it came to be. For some use cases, this is sufficient, but in other situations, more care is needed in drawing conclusions from geotagged social media data. Below, we discuss some salient issues for researchers in using geotagged data.

AVAILABILITY OF COORDINATE GEOTAGS IS DECREASING

Publicly visible geotagged social media may seem like an endless source of rich data. However, as we saw in Study 1, Twitter has fewer coordinate geotags available, and even Flickr geotags may be plateauing. Additionally, our participants in Study 3 geotagged primarily on Facebook, Instagram, and Foursquare, but these services do not publish public "firehose" APIs for gathering data.

Furthermore, the geotags that are still present are getting stranger: job posting bots, weather and sports bots, deleted accounts, and other accounts are creating a growing fraction of all public geotagged tweets.

As a result, it is not clear how much more research can be done with coordinate geotags. In addition, for the coordinate geotags that are available, it is important to dig in and filter out whatever spam may be prevalent.

An alternative is that researchers may need to become more comfortable dealing with placetags. For example, it is important to avoid the center-of-rectangle problem mentioned in the previous section. The semantics of a placetag may also vary by application. A few general principles include realizing that the data available at the building or neighborhood level will likely be much smaller than the data at the city level, and the data at the point level smaller still.

GEOTAG PROVENANCE AFFECTS RESEARCH METHODS

It is important to treat geotagging as a performative act, not a passive one. As Study 2 showed, most people consciously decide to geotag each time; as Study 3 showed, people use geotags mostly to show off where they have been, keep family and friends updated, and occasionally coordinate with friends or save a place for later. The content of their tags often reflects vacations or meals at restaurants. As a result, it is a rich data set to study where people go on vacation, eat out, or have places that they want to save for later. However, it does not seem to be a rich data set to study users' everyday lives.

As an example of a use case where the provenance of these tweets matters, we point to [87]. In this work, the authors tried to find users' home addresses given a sample of their geotagged tweets. If tweets represented a random sample of places the user has been, this would be trivial, because most tweets would occur at users' homes. However, they found that this was impossible for about 15% of Twitter users, because they recently moved, never tweeted at home, or had other complicated use cases. They reference Krumm [47], who previously attempted to find home addresses based on GPS sensors on cars. Naturally, with the same methods, the GPS on the car worked much better, because those readings come from a passive and automatic sensor, not a performative act.

But the fact that geotags come from unusual occasions doesn't only limit research; it can add extra context to a post. For example, many geotagged tweets come from Untappd, an app for beer aficionados to share their experiences of fine beers. While Untappd tweets tell us less about the general public's day to day movement, they tell us more about local beer loving communities and, potentially, the sociability of a place. Likewise, the fact that people often geotag on vacation or while out to dinner may provide clues to activity recognition and computer vision algorithms. We encourage researchers to find questions that can take advantage of the rich variety of sources that geotagged tweets provide.

3.10 CONCLUSION

While checking in in location-based social networks has become a widely researched topic, motivations behind geotagging in other social media have not been as fully analyzed. We investigated if people geotagged their social media posts for the same reasons that they checked in on LBSNs, and for the most part found that they do. Geotags are postcards, not ticket stubs; conscious choices, not byproducts. People geotag their social media posts to show off where they are or communicate with family and friends. They geotag in faraway and rare places. However, in our research we found reasons, especially recently, why researchers and developers must be careful. Researchers should be aware that counts of coordinate geotags are shrinking and specific types of spam are rising, while developers should show their users clearly what they are posting, avoid

sticky geotagging settings, and prioritize placetags over coordinates.

Returning to our research questions, these findings show that geotagged social media posts are likely to be useful to show glimpses into a world. Because they are provided by a small subset of the users of any particular social network (who are in turn a small subset of a population), they are less useful for demographics or other statistics about a population as a whole. This suggests that they will be likely to be a valuable resource for travelers more than for city planners or other officials.

4 WHAT TRAVELERS WANT

After surveying what we know about social media, I turn now to the question of how it can help travelers. But before answering that, it is important to understand these new tourists' motives. In this chapter, I address the question, "What do these new travelers mean by 'get a feel for the city'?" How can we operationalize this vague concept to help them find what they're looking for?

In this chapter, I detail two studies: first interviews with 14 creative tourists, then surveys with 490 more creative tourists, to find out what they are looking for. The model that emerged has five parts: safety, convenience, liveliness, aesthetic appeal, and the ability to live like locals. This further informed the design of the Neighborhood Guides application and future research, as described in chapters 5, 6, and 7.

4.1 INTRODUCTION AND BACKGROUND

The sharing economy has revolutionized the tourism industry. Apps like Airbnb, Couchsurfing, and HomeAway allow people to match up with locals and stay at their homes. Airbnb now boasts over 60 million users staying at over 2 million listings [1], Couchsurfing has over 10 million users [2], and HomeAway has 1.2 million listings [3]. Mobile and web technology, user reviews, and social media have combined to enable people to stay in tons of new places.

At the same time, a growing body of research suggests that modern urban travelers want to travel in new ways: they want to experience "everyday life" in a city and "do what the locals do." Unlike the sun-and-sand tourists of two generations ago or the cultural-site-visiting tourists of last generation, today's tourists want to curate and create their own experience. Choosing unique lodging is one way they can do so. However, travelers often don't know what neighborhoods they might like to stay in. The plentiful lodging around the city gives travelers an overwhelming array of choice, and without local knowledge of a place, travelers don't know where to start looking.

Tools that are available to address these information needs all fall short. Traditional guidebooks from Fodor's, Frommer's, and Lonely Planet give people information about those central tourist districts and sights to see. Yelp and Foursquare give people information about the businesses, the bars and restaurants and locksmiths, in an area, but travelers cannot easily understand how the whole neighborhood feels just from that. Cities gather statistics - and indeed, are releasing open data more than ever before - but numbers also fail to convey a neighborhood's culture. Finally, occasionally travelers can learn local vernacular descriptions, but these are often shallow. For example, "Lawrenceville is the cool neighborhood" or "South Side is the party neighborhood."

In order to design tourist guides of the future to help people renting through the sharing economy, we conducted two studies of travelers. We first interviewed 24 people who have recently traveled or moved to another city. We then surveyed 490 people (98 recruited via social media and

392 users of Airbnb) to better understand what they want to know about neighborhoods. We developed a five-dimension model of what users want in neighborhood search: safety, location convenience, liveliness, the ability to live like locals, and aesthetic appeal. This model, and the understanding built through the qualitative and quantitative studies, will help us design modern neighborhood guides to help these travelers.

4.2 STUDY 1: INTERVIEWS

For this research, we employed a mixed methods approach, gaining qualitative insights from interactive interviews, which we adjusted and confirmed with quantitative data from surveys (described later as Study 2). We began with interviews of 24 participants.

In these interviews, we focused on the following research questions:

- What do people want to know about neighborhoods when they're traveling?
- What do people want to know about neighborhoods when they're moving? (Note: we later removed the movers from our sample to focus on travelers.)
- What do people wish travelers and movers knew about their neighborhood?

We recruited 17 participants in Pittsburgh who all recently traveled or moved by posting our study on Reddit, Craigslist, and Facebook. We asked them to describe their search process and their experience finding a neighborhood to stay or live. We then showed them printed pages about the neighborhoods they moved from or traveled to: popular Twitter words that are more common in that neighborhood than others, Flickr photos obtained by searching the neighborhood names on Twitter, the top 10 most popular venues on Yelp, and market research and statistics from the Tapestry guide produced by GIS company ESRI [1]. We asked them to create two one-page guides (one for the neighborhood they moved from/traveled to, and one for the neighborhood they currently live in) by cutting and taping these materials, and writing or drawing in anything that was missing. This was meant as an elicitation exercise to get them thinking about these neighborhoods in more depth. Our university's Institutional Review Board approved this study.

After these 17 interviews, of which 7 involved recent travelers and 10 involved recent movers, we realized one recurring issue: location guides often matter more to travelers than movers. Movers care about many factors in addition to the neighborhood: the house or apartment itself, the cost of rent or a mortgage, the proximity to an existing job, and the local schools. They also tend to have much more time to investigate neighborhoods before moving there. Some of their concerns still echo the travelers' concerns, so we retained their data, but we reoriented the project to focus on travelers.

We recruited seven more recent travelers in San Francisco, bringing the total to 24 participants. For this second group, we did the same interview, but focused more on factors that seemed relevant in the first one: safety, liveliness, diversity, and aesthetics. We also only recruited travelers for the second group. We did not bring the printouts or ask participants to create flyers like we did for the first group, because the extra work did not provide much more value or insight.

[1] www.esri.com/data/esri_data/tapestry

We will refer to the original 17 interviewees as A1-A17, and the next seven as B1-B7. Basic information about these participants is included in Table 4.1. Interviews were conducted in cafes or other public places near them for convenience and to get them thinking about their neighborhoods. B5 and B6, a dating couple, interviewed together; all the rest were done separately. Our participants are mostly in their 20s and 30s, which are also the age groups most likely to try a home-sharing site like Airbnb or Couchsurfing, according to a 2016 Pew Internet report [81].

Because the interviews occurred in public places, we could not record them, but we took notes to capture important points as well as possible. After finishing each batch of interviews, we analyzed the data iteratively, using an open coding approach to allow insights to emerge from the data. In this method, as described by Strauss and Corbin [83], a researcher iteratively reads through notes, such as statements or opinions from participants, and applies codes to describe higher-level themes relating to these notes. After completing all the notes, the researcher re-reads the notes and re-codes them until they reach convergence. In this way, the higher-level themes can emerge from the data.

These interviews revealed a lot about this group's travel and moving motivations, what they hope to learn about neighborhoods, and where they decide to stay, as well as a few interesting tensions that arise when they make those decisions. We built a six-part model of tourist information search, which we then refined and verified with our next study.

4.2.1 TOURIST NEIGHBORHOOD SEARCH MODEL 1.0

After coding interviewees' responses, we came up with a six-dimension model that it seemed users would search on. We describe it here for clarity, though we will later introduce a more validated model that we believe to be more accurate. To reduce redunancy, we also offer more details about these dimensions by combining the results of these interviews with the results of our later survey.

Our participants wanted to search for neighborhoods along the following dimensions:

1. Safety: how easy it is to avoid crime or other trouble
2. Diversity: how diverse the people are who live there
3. Walkability: how easy it is to get around by walking
4. Aesthetic Appeal: how nice the neighborhood looks
5. Third Places: how wide and deep a selection of bars, cafes, and other social places exists in the neighborhood. (We use the term "third places" as introduced by Oldenburg [66].)
6. Authenticity: participants defined this in many different ways, often invoking a duality with "touristiness" (an "authentic" and desirable place was one that was not "touristy.") We define authenticity more precisely when discussing the later model.

4.3 STUDY 2: SURVEY TO VALIDATE AND CLARIFY MODEL

To refine our model after Study 1, we conducted a survey. We wanted to know:

	Age	Occupation		To/from where?
A1	20-29	PhD Student	Travel	New York City
A2	20-29	Social Worker	Move	Within Pittsburgh
A3	20-29	PhD Student	Move	Palo Alto, California
A4	30-39	System Administrator	Move	Slippery Rock, Pennsylvania
A5	20-29	Master's Student	Move	Mountain View, California
A6	20-29	Master's Student	Travel	London and France
A7	30-39	IT/Network Engineering	Move	Philadelphia
A8	30-39	PhD Student	Move	Maryland
A9	20-29	Social Media Manager	Travel	Philadelphia and Toronto
A10	20-29	Unknown	Move	Within Pittsburgh
A11	20-29	PhD Student	Travel	Zurich, Switzerland
A12	20-29	Master's Student	Move	Bangalore, India
A13	20-29	Lawyer	Move	San Juan, Puerto Rico
A14	20-29	Research Assistant	Travel	Montreal
A15	50-59	Museum Exec. Assistant	Move	Philadelphia
A16	20-29	Unemployed	Travel	Asheville, North Carolina
A17	20-29	Institutional Researcher	Travel	Philadelphia
B1	30-39	Assistant Professor	Travel	Norway and Houston
B2	20-29	Unknown	Travel	Oxford, London, and Mumbai
B3	30-39	Unknown	Travel	Lake Tahoe
B4	20-29	Unemployed	Travel	Scottsdale, AZ; Florida Keys
B5	20-29	Musician and teacher	Travel	Seattle and Aspen, CO
B6	20-29	Travel journalist	Travel	Seattle, Aspen, New Orleans, London
B7	20-29	Tax consultant	Travel	Merida, Mexico

Table 4.1: Participants in Study 1 (Interviews). Participants in group A lived in Pittsburgh; group B lived in or near San Francisco. Each was asked about a recent time they traveled or moved (group B was only asked about travel).

- Is our six-dimension model of tourist neighborhood search complete, or have we missed any important dimensions?
- Are all six dimensions necessary?
- How important is each dimension?

The survey asked participants which of our six dimensions was the most important thing when they are finding a place to stay (with an option for "other"), then asked them more nuanced questions about the relative importance of each one. These questions were taken from key points people brought up during the interviews; for each of our six dimensions, we created 2-4 questions based on that dimension. The survey questions are shown in Table 4.2.

We recruited participants in two batches. In the first batch, we recruited on Facebook, Twitter, Reddit, Craigslist, and Slack, and through a participant pool at our university. This survey was approved by our Institutional Review Board. The survey took about 10 minutes, and participants were entered into a drawing for one of five $50 Amazon.com gift cards. We received 98 responses. For the second batch, our collaborators at Airbnb, Inc., sent the survey to some of their users and received 392 responses.

4.4 Results

4.4.1 Travelers use social or budget heuristics if possible

A number of conditions may cause travelers to do very little research before choosing where to stay. If someone already has a place to stay, they will likely take that. B2 described this as a "bird in the hand" situation, and said it occurred a lot when Couchsurfing: finding a local who's willing to host him for free can be difficult, so he will usually accept, regardless of circumstances.

If a traveler has social or other constraints, such as friends or family to visit or an event to attend, they usually consider tourism secondary and stay somewhere nice near that constraint. B5 and B6 described going to the X Games, an extreme sports event, in Aspen, Colorado: they spent most of their time watching events, so they simply wanted to stay near the games. Similarly, B4 described visiting Scottsdale, Arizona, on per sonal business, which led to him staying in the Fashion Square district. He found it rather unpleasant, and had trouble getting around, but he needed to be near there.

Finally, budget constraints would often short-circuit the lodging search. B5 and B6 described another trip, when they went to Seattle but wanted to pick the cheapest lodging possible. This ended up being the Green Tortoise Hostel downtown, and since they had stayed in another Green Tortoise elsewhere, they decided it would work. B3 also described a road trip where he simply looked up a place to stay while on the road each day, only wanting something simple, clean, and cheap.

4.4.2 If no heuristics are available, people try to optimize along five dimensions.

Most of the participants in Study 1 described at least some trips where they did not use any of these heuristics, and instead wanted to satisfy six different dimensions: Safety, Diversity,

Dimension		Question text
	1	Which of the following is the most important to you when finding a neighborhood to stay in when you travel? (Safety; Diversity of people there; Walkability; Aesthetic appeal; Cafes, bars, and social spaces; Authenticity; Other (enter your answer))
Safety	2	How concerned are you with safety when you travel?
	3	How influential is an area's crime rate in deciding where you will stay?
Diversity	4	When you travel, how desirable is it for people who live in the area you're staying in to be diverse?
	5	How often do you go to places where lots of different people interact?
	6	Would you rather stay in an area that is "up and coming" or an area that is "established"?
Walkability	7	How important is it to be able to get around by walking when you travel?
	8	When you travel, how desirable is it to be in an urban place with lots of activity?
	9	How important is it to be able to get around with public transit when you travel?
	10	How often do you have a car (whether you drive it to your destination or rent it there) when you travel?
Aesthetics	11	When you travel, how important is it that the neighborhood you stay in looks nice?
	12	How influential is the "look" of a neighborhood in choosing where you want to stay?
3rd places	13	How important is it that the neighborhood you stay in has great bars, cafes, or other social spaces?
	14	When you travel, how often do you speculate what life would be like if you lived there?
	15	When you travel, how often do you try to do what the locals do?
Authenticity	16	When you travel, how desirable is it to stay in an area that caters to tourists?
	17	How important is it to you to find "off the beaten path" places when you travel?
	18	How important is it that the neighborhood you stay in is also a functional neighborhood for people who live there?

Table 4.2: Survey questions for Study 2. All questions after question 1 were presented in random order and had 5-point Likert scale responses. "How desirable" questions (4, 8, 16) had responses from "Very undesirable" to "Very desirable." "How Important," "How Influential," and "How Concerned" questions (2, 3, 7, 9, 11, 12, 13, 17, 18) had "Not at all/Slightly/Somewhat/Very/Extremely" responses. "How often" questions (5, 10, 14, 15) had "Never", "Almost never", "Occasionally", "Almost every time", "Every time" responses. Question 6 had "Much prefer up and coming", "Somewhat prefer up and coming", "Neutral", "Somewhat prefer established", "Much prefer established" responses.

Aesthetic Appeal, Walkability, Third Places, and Authenticity. Figure 4.1 shows how many people selected each of these six as their most important dimension.

After Study 2, we adjusted the model and instead found five dimensions. We did this by performing factor analysis on our survey questions. Five eigenvalues of the correlation matrix were greater than 1 in both datasets, so after performing factor analysis with 5 factors, loadings are shown in Table 4.3.

We will explain these dimensions in the next section.

4.4.3 CURRENT NEIGHBORHOOD SEARCH TOOLS ARE INADEQUATE

Given that these five dimensions matter in different ways for different searchers, how do travelers search for neighborhoods now?

The primary search method participants said they used was to ask friends and family. If people visited friends, like B2 in Albuquerque and Portland, they can do this directly; otherwise, like A9, they would ask friends beforehand what were interesting and fun neighborhoods.

Online research was also widely used, often as simply as searching Google for "things to do in London" or "London off the beaten path" (B6). B7 lamented, though, that this kind of searching can turn the usually-fun process of traveling into work.

Because searching was so labor intensive, some people who did not have any pre-existing heuristics (as described above) tried to create their own heuristics. A11 would search for the "queerest neighborhood" in a given city, as she did when she visited Zurich. This was not in order to find particular sites there (Zurich's queerest neighborhood featured one main gay bar and one main sex shop, neither of which she visited), but just because she found that she would often like the kind of people she met there. Similarly, B1 searched for the best coffee shops, not because she would spend most of her time there, but because she usually likes neighborhoods that have good coffee shops. B2 would read books about a place, like Gregory David Roberts's novel *Shantaram* before visiting Mumbai, or Maya Angelou before visiting San Francisco, in order to recognize places they mentioned.

4.5 TOURIST NEIGHBORHOOD SEARCH MODEL 2.0

Our refined model of tourist information search has five dimensions: the primary dimensions of Safety and Location Convenience, and the secondary dimensions of Living Like Locals, Liveliness, and Aesthetic Appeal. Figure 4.2 shows which parts of our Tourist Information Search Model 1.0 turned into each dimension in this Model 2.0.

4.5.1 DIMENSION 1: SAFETY

Everyone wanted to be safe. Safety was a dimension that came out of our interviews, and our surveys verified its importance. In both the general public and the Airbnb user populations, people most commonly chose safety as the most important factor (38 of 98 public respondents and 153 of 392 Airbnb respondents).

	Factors				
Question	1	2	3	4	5
2 (importance of safety)	0.85				
3 (crime rate)	0.85				
4 (diverse people)			0.32		
5 (where different people interact)			0.44		0.35
6 (“up and coming” vs. “established”)					
7 (get around by walking)		0.62			
8 (urban place)		0.33			0.47
9 (get around with public transit)		0.73			
10 (have a car)		-0.67			
11 (looks nice)	0.31			0.59	
12 (the look of the neighborhood)				0.93	
13 (bars and cafes)					0.60
14 (speculate what it’s like to live there)			0.35		
15 (do what locals do)			0.62		
16 (caters to tourists)					
17 (off the beaten path)			0.62		
18 (functional neighborhood)			0.33		

Table 4.3: Factor loadings on survey questions, Airbnb data set. (findings on the general public data set were similar.) Loadings <0.3 are omitted. Factors 1 through 5 became safety, location convenience, living like locals, aesthetic appeal, and liveliness, respectively.

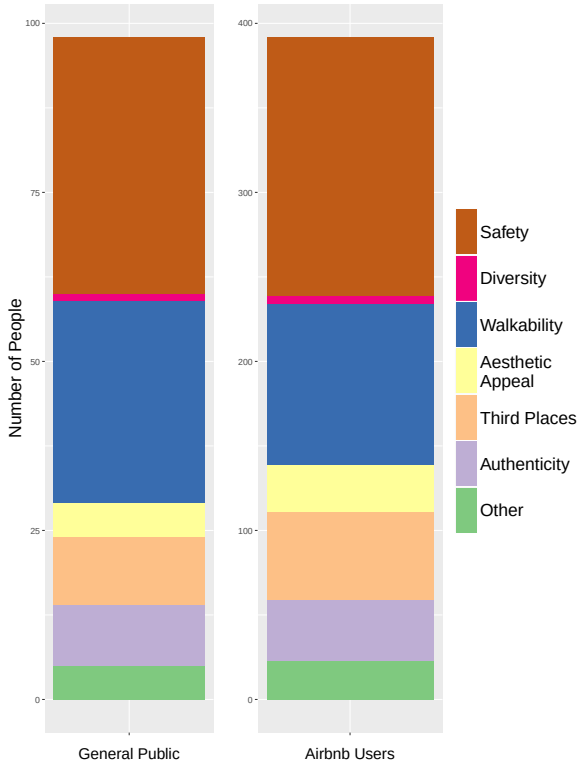


Figure 4.1: The most important dimensions of choosing a neighborhood to stay in, according to our respondents (general public on left, Airbnb users on right).

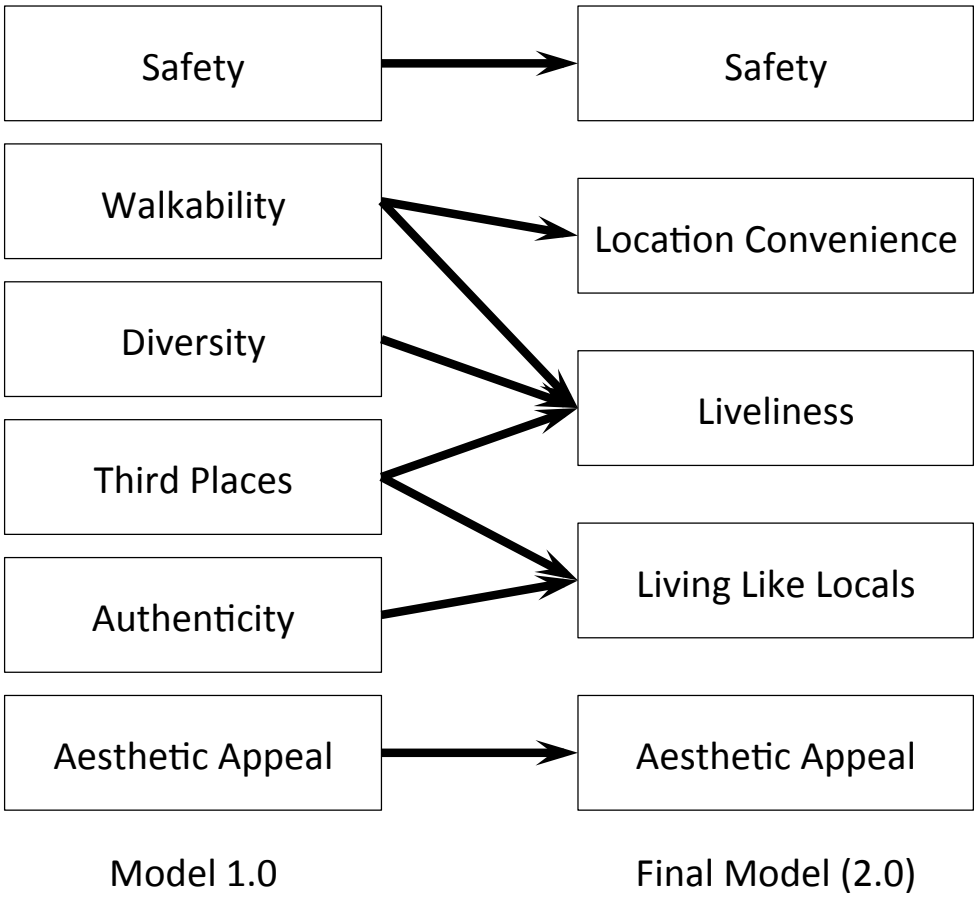


Figure 4.2: Our original model and final model after Study 2.

The meaning of safety varied slightly depending on location; usually it included crime, but A1, A15, A17, and B4 all mentioned fear of bedbugs when traveling to New York. People's interpretations of "safety" varied, too, depending on the type of crime and individual thresholds. A2 mentioned the distinction between "drug deal gone bad" vs. "random" crime, A5 looked up murder rates rather than crime rates because those were the crimes that mattered the most, and B4 didn't really care about most crime, except, "I just don't want to get shot at." On the other hand, A6 rerouted a whole itinerary through France after hearing that Marseille, one of their planned destinations, was "unsafe."

When asked if safety was always an upside, many participants declined. A6 described spending one night in Churchill Gardens, a posh part of London, but then moving on to somewhat simpler Clerkenwell. Often the safest spaces are also the most expensive, and because they are so expensive, only a homogenous set of wealthy people can live there.

4.5.2 Dimension 2: Location Convenience

We define the "location convenience" of a place as how easy it is for a traveler staying at that place to get to everywhere they want to go. This concept emerged from our survey results; we had missed it in the interviews, in favor of the concept of "walkability." Walkability is part of location convenience: if a place is easy to walk around, then it will be easy to get to attractions and daily necessities. Especially in a foreign place, where one might not have a car or understand local public transport, walking to something is often the easiest way to get there.

However, location convenience extends beyond walkability, depending on the local transportation options and the traveler's trip. A1 and B1 both talked about the extensive subway in New York City; A1 noted that she did not feel compelled to stay in central Midtown as long as she was near a subway, while B1 preferred taking subways to walking because it was an interesting experience in itself. Perceptions of location convenience change with mode of transportation and circumstance, too. B5 and B6 described going to Aspen, Colorado together to see the Winter X Games extreme sports competition. Three areas, Aspen, Buttermilk, and Snowmass, had lodging options, but the only transport between them and the X Games site was by bus, and Buttermilk and Snowmass required an extra bus ride to get to Aspen. B5 and B6 focused their search on Aspen itself, because it would require one bus ride per day instead of two. Naturally, in a less crowded time of the year, Buttermilk and Snowmass would be more location convenient.

Location convenience for travelers is related to location efficiency for people living in a place [33]. Location-efficient places are dense places with good public transit access, where residents will be less likely to need to own a car. Location efficiency and location convenience are not exactly the same, as travelers have a different set of priorities (for example, getting to a tourist destination rather than a job, or a restaurant rather than a grocery store), but they are quite similar.

We did not ask specifically about location convenience in Study 2; its inclusion came out of the "other" responses. Of the 23 "other" responses from the Airbnb group, 13 were about location convenience, as were 2 of the 4 "other" responses in the general public group. Also, as walkability was the second most common choice as the most important dimension (see Figure 4.1), we assume that location convenience would be the second most common choice if we had included

it instead.

4.5.3 DIMENSION 3: LIVELINESS

Our participants mostly appreciated lively places. This intuitively makes sense, as they traveled to cities; the trips they discussed are not "sun and sand" getaway trips. Lively places have a number of advantages: there are usually businesses nearby, so it is easy to accomplish daily tasks; there are fewer safety concerns; they are often interesting in their own right due to markets or street performers. "Liveliness" encompasses what we previously referred to as "diversity" because the street is an equalizing place, as Jacobs writes [35]. All kinds of people can meet on a street; as A1 and B1 said, their favorite places allow "room for everyone."

Liveliness thus includes some of the components of diversity, some components of walkability, and some components of third places. Some participants offered examples of lively places they enjoyed: B1 enjoyed train stations, while she and B6 both brought up markets. Others described liveliness in their own words: "being in the middle of stuff" (B4), "having stuff around" (A15), or being "where everything is" (A9). Liveliness makes traveling more pleasant and enables serendipitous encounters too. B6 talked about visiting New Orleans and stumbling across parades put together by local Native American groups, which she unexpectedly enjoyed. Similarly, B7 described how she preferred staying in the residential Itzimna neighborhood over the center of Merida, because she enjoyed her 20-30 minute walk to the center every day and the chance encounters it brings.

4.5.4 DIMENSION 4: LIVING LIKE LOCALS

This dimension may be the most complex, as it includes two related ideas: the authenticity of a place and the opportunity to experience everyday life. Due to its complexity, we will discuss it in two parts.

AUTHENTICITY

This dimension represents how closely people can simulate a "normal", non-traveling life there, and approximates notions of "authenticity." Many participants expressed desires for an "authentic" "non-touristy" place. Clearly, "touristy" places have some disadvantages: they are expensive (B6 gave the example of paying £39 to see the Crown Jewels in London) and often people act differently there (B7 described feeling like she "had a dollar sign on her forehead" in the tourist beaches of Cancun). But those inconveniences do not explain the intensity of the desire to be "not a tourist" (or even "the anti-tourist", as A9 described himself). Furthermore, some people appreciated touristy places, for practical reasons: B7 noted that not speaking Spanish limited her experience in Mexico, and A6 described how she would search for a place that's not the #1 tourist destination but also not completely local, due to language issues.

To understand this touristiness tension, it is useful to review previous work about authenticity in tourist places. Early work located all spaces on a 6-stage scale from front-stage (purely for show) to backstage (fully authentic) and predicted that all tourists would seek authenticity [56]. Later work added more nuance, describing the "authenticity" of an experience in nine subtypes depending on how authentic the place was, how authentic the people were, and whether the visitor

put importance on the authenticity of the people or the place, both, or neither [68]. Furthermore, the authenticity of an experience may be best explained as existential authenticity, or the personal resonance with that experience. Existential authenticity has two forms: intra-personal (discovering and being true to oneself) and inter-personal (having a real connection to others) [95].

For example, different people may enjoy a trip to the Van Gogh Museum in Amsterdam for many reasons. They may appreciate seeing the original Sunflowers (objective authenticity) or seeing the official, definitive collection of Van Gogh's art (constructive authenticity). They may enjoy a stirring resonance with Van Gogh's masterful brushstrokes, or the ability to discuss these paintings with their fellow tourists (existential authenticity, intra- and inter-personal respectively). They might get useful information from the docents in an official, front-stage capacity, or they might get a docent to reveal little-known backstage stories about working at the museum. Finally, afterwards, they may stay in the enclavic tourist "bubble" of the Museumplein outside, or they may head to a more heterogeneous neighborhood, as described in [22]. Each of these experiences may be regarded by one person as "authentic" and by another as "touristy."

THIRD PLACES

Another recurring theme was described as "taking in the city life" (B1), seeing "what people actually do here" (A9), "kind of get[ting] a feel" of the city (A6), and even "play[ing] the game of, what if we lived here" (A17). This echoes a trend towards travelers using the everyday as a way to create their experience, for the travel experience to be less about what they are consuming and more about what they are becoming [57]. Most participants (except A16) were not traveling in order to find a place to move to, but they still enjoyed pretending to do so.

When pressed, though, interviewees did not actually want their travel experiences to be about the real "everyday." Everyday life involves work, chores, and errands that most people do not enjoy, wherever they are. For example, asked if she would be interested to see everyday life in the Financial District of San Francisco, B1 replied no, the Financial District isn't the kind of "everyday" she's looking for (though clearly it is an integral part of many people's everyday lives). Instead, participants wanted to experience an "ideal everyday," which involved two recurring subthemes: relaxation and third places.

Relaxation is self-explanatory: travelers, usually on leisure trips, preferred a slow-paced day with few responsibilities to a quick, busy day. A1 appreciated a relaxing or "chill" environment, as did B1, who elaborated that, as a busy professor, she often doesn't get a chance to do the "everyday" things that are part of this ideal day. She gave an example of buying a birthday card for a friend: she plans to do this on an upcoming trip to visit friends, just because that will be the only time she has to do it.

Third places, such as bars, cafes, and bookstores as described in [66], are also a key part of this "ideal everyday." Many participants described local venues they loved: a coffeeshop and a taqueria (A13), cafes where one can see friends sitting outside (A14), cafes and dive bars (B4). B1 went as far as to suggest that she would travel to a place based on where the best coffee shops were. Because third places tend to be neutral, accessible, status-leveling places, travelers

appreciate them. Stepping into everyday life in another place involves adjustments, and these third places give travelers a way to recharge.

After our factor analysis of the survey results (see Table 4.3, we realized that the concepts of authenticity and third places, which we had assumed to be separate, really reflected the same underlying emotion: travelers want to be able to be part of a different place, to experience a different life instead of just seeing some different things. Therefore, we combined these two dimensions into one overall theme of "living like locals."

4.5.5 Dimension 5: Aesthetic Appeal

Aesthetic appeal in many forms is one of the main incentives for people to travel, and one of the main influences on the overall feeling of a trip. By "aesthetic appeal", we are referring to anything about the senses: participants mentioned visual, auditory, and gustatory appeal particularly, and occasionally smell. Some preferences were universal, such as enjoyment of nature and avoidance of loud places while sleeping. Others were personal: A10 described her neighborhood as a burgeoning urban agricultural area, while A4 described the city of Pittsburgh as a "concrete jungle." Many participants described suburbs as "boring", but A7 described one suburb as his "perspective of what country living should be."

4.6 Discussion and Design Opportunities

This model provides both deep insight into how new urban tourists search for neighborhoods to stay in, and shows ways forward to make guides to better serve them. In this section, we will discuss some design recommendations for these systems.

4.6.1 Focus on areas, not points

Current tourist guides focus on individual places. Tourist sites, restaurants, shows, shops, and bars are all approached as if one had a perfectly rational choice between them. Searching for Indian food on Yelp results in a list of nearby places, and a traveler can just pick the highest rated option. Plenty of the research we reviewed in this paper focuses on venue recommendation as well.

But modern urban tourists, especially, want to know about interesting areas, not just interesting points. They enjoy traveling by experiencing the ideal everyday life of an area, relaxing in cafes, browsing shops, reading a book in a neighborhood park. Reading about statistics or lists of top 10 restaurants will not help them find the neighborhood they would like.

As a result, we see an opportunity for a higher level of abstraction. One promising avenue is comparison to neighborhoods in cities that they know. This is similar to work that has been done both in research [51] and in popular culture [74]. Because people already know what neighborhoods in their own city are like, this can give them an easy way to understand neighborhoods in a new city.

4.6.2 SAFETY AND LOCATION CONVENIENCE MAKE A LODGING OPTION GOOD ENOUGH

Travelers need to get around, and they need to be safe. These were almost universal requirements from the participants that we spoke to, and the survey responses confirmed their importance. Fortunately, these seem to be easy dimensions to address in an application. Safety can be addressed using public crime data that many cities are already publishing. Location convenience is a bit more difficult, but services like Walkscore[2] and Mapnificent[3] are integrating walking and transit to show how hard it really is to get around from a given place.

4.6.3 LIVELINESS, LIVING LIKE LOCALS, AND AESTHETICS MAKE IT GREAT

After a place has passed a bar of "safe enough" and "convenient enough", modern tourists are looking for a lifestyle. They want to experience the "ideal everyday" life in a place, they want a place that's exciting to walk around, they want a place that looks nice. Unfortunately, including these in an application is not quite as easy. However, social media offers some promising possibilities.

First, liveliness can be approximated using residential density and location reviews on sites like Yelp and Foursquare. Denser places will be more lively and likely have more places, and more reviews. Some way to look into the words in those reviews would help too: if there are tons of bars that people are praising for their wild parties, that will give a neighborhood a different feel than if it's full of quiet restaurants or family-friendly cafes.

Second, the ability to live like locals could be mined from Yelp, Twitter and/or Flickr. As we described previously, "living like locals" involves at least two parts: the third places in the neighborhood, and the neighborhood's authenticity. Therefore, the reviews of third places in Yelp can help. Authenticity, though, is (as we reviewed) an ill-defined concept. Some part of authenticity involves how many tourists and how many locals go there, which could be shown via a tool like Eric Fischer's "Tourists and Locals" maps[4]. But this is not enough; authenticity is individual. Perhaps an aggregation the tweets in an area could show something about what people talk about in that area, and someone could decide if what people talk about is "authentic" or not.

Third, aesthetic appeal might be able to be described using Flickr photos. There are already 100 million photos publicly available through the YFCC100M dataset [89], and more through their public API. More importantly, these photos have computer vision-based automatic tags, so one can easily see which visual features (such as "car" or "dog" or "beach") appear frequently in a neighborhood. Summarizing a neighborhood's photo autotags could potentially help travelers visually explore neighborhoods easily.

4.7 CONCLUSION

New urban tourists want to stay in interesting residential neighborhoods and spend time "wandering about", "taking in the city", and "getting among the people" [6], and our participants echoed these previous findings. To operationalize these desires, we conducted a series of interviews with

[2]https://www.walkscore.com
[3]http://www.mapnificent.net/
[4]https://www.mapbox.com/labs/twitter-gnip/locals/

14 tourists (and 10 recent movers) and surveyed 490 travelers. We built the 5-dimensional model of creative tourist information search: tourists want safety, location convenience, aesthetic appeal, liveliness, and the ability to live like locals. In the next chapters I will explain how I used this model to build a prototype neighborhood guide and run user studies based on it.

5 NEIGHBORHOOD GUIDES PROTOTYPE

In this chapter I will describe the prototype neighborhood guide application that I used as a probe for future studies, and what I learned from its initial evaluation.

This neighborhood guide was not intended as a finished product, but rather as a probe to elicit reactions from participants to better understand what they need and want when traveling, and to see which side of the city is reflected by the social media that is posted there.

5.1 OVERVIEW

The Neighborhood Guides application was conceptualized as a website that travelers could visit while planning a trip. It contains information about each neighborhood they might consider staying in.

The information included is based on the five-dimensional model of creative tourist information search we built in the previous chapter. Each dimension is represented by a data source from social media or another publicly available source. Specifically, to address the dimension of safety, we use crime statistics from city open data portals; to address convenience, we include information from Walkscore[1]; to address aesthetic appeal, we include photos from Flickr; for liveliness, counts of venues of different kinds from Foursquare; for the ability to live like locals, common words from Twitter. These data sources are shown in Table 5.1.

[1] http://www.walkscore.com

Model Dimension	Data Source	Information Derived
Safety	City open data portals	Crime counts by neighborhood, by type
Convenience	Walkscore	Walk score, Bike score, Transit score
Aesthetic Appeal	Flickr	Photos, sorted by autotags
Liveliness	Foursquare	Counts of venues, by type
Living Like Locals	Twitter	Common words and context

Table 5.1: Dimensions of the Creative Tourist information search model and corresponding data sources we used for our Neighborhood Guides prototype. More detail about how we derived information from these data sources is provided throughout the rest of this chapter.

5.1.1 CITY AND NEIGHBORHOOD CHOICES

One important choice we had to make before beginning was which cities and neighborhoods to include. We chose Pittsburgh and San Francisco because I was very familiar with them both, and Chicago and Houston because they were reasonably large cities with plenty of data available.

For each city, we used their open data portal to find neighborhood boundary data. The portals included the Western Pennsylvania Regional Data Center[2] for Pittsburgh, SF OpenData[3], City of Chicago Data Portal[4], and Houston Data Portal[5]. If they had multiple neighborhood boundary data sets, we selected the one that seemed to be the most canonical or widely used.

One choice we made was to only use officially designated neighborhood boundaries. Previous projects like LiveHoods [18] and Hoodsquare [97] have found success using Foursquare checkins to define socially-connected neighborhoods, but we consciously chose not to use these metrics for a number of reasons. Most prominently, we wanted them to have names that connected with external data sources and people's intuitions. For example, we wanted users to be learning about "Squirrel Hill" or "The Mission", not "LiveHood #34". This would aid in the external validity of any information they learn while using the site. In addition, using official boundaries deflects responsibility from us. People identify strongly with their neighborhoods and have strong opinions about where their neighborhood ends and the next neighborhood begins; we wanted them not to be distracted by these distinctions in our studies.

This site had a long vertical layout; users could see everything in Figures 5.1, 5.2, and 5.3 by scrolling up or down.

5.2 INFORMATION ABOUT EACH NEIGHBORHOOD

As mentioned previously, we used publicly available data from open data portals to describe where the boundaries of each neighborhood lie. We then gathered information about each neighborhood as follows.

5.2.1 SAFETY: CRIME STATISTICS

Many cities now release easily-parseable crime reports, so we gathered one of these for each city. In order to include a fair sample of each city without needing to correct for seasonal trends, we took data from all of 2015, which was the latest year fully available. Most cities did not have this data available by neighborhood, so we took a list of all crimes in each area and split them into neighborhoods. We also categorized them using the FBI UCR Offense Definitions [63] into Part I (more serious) and Part II (less serious) offenses, to help with the suggestions from participants in chapter 4 that different types of crime may or may not worry them. We scaled these values by the population of each neighborhood. Eventually, we had three values for each neighborhood: Part I crime, Part II crime, and Total crime.

[2] http://www.wprdc.org
[3] https://data.sfgov.org
[4] https://data.cityofchicago.org
[5] http://data.ohouston.org

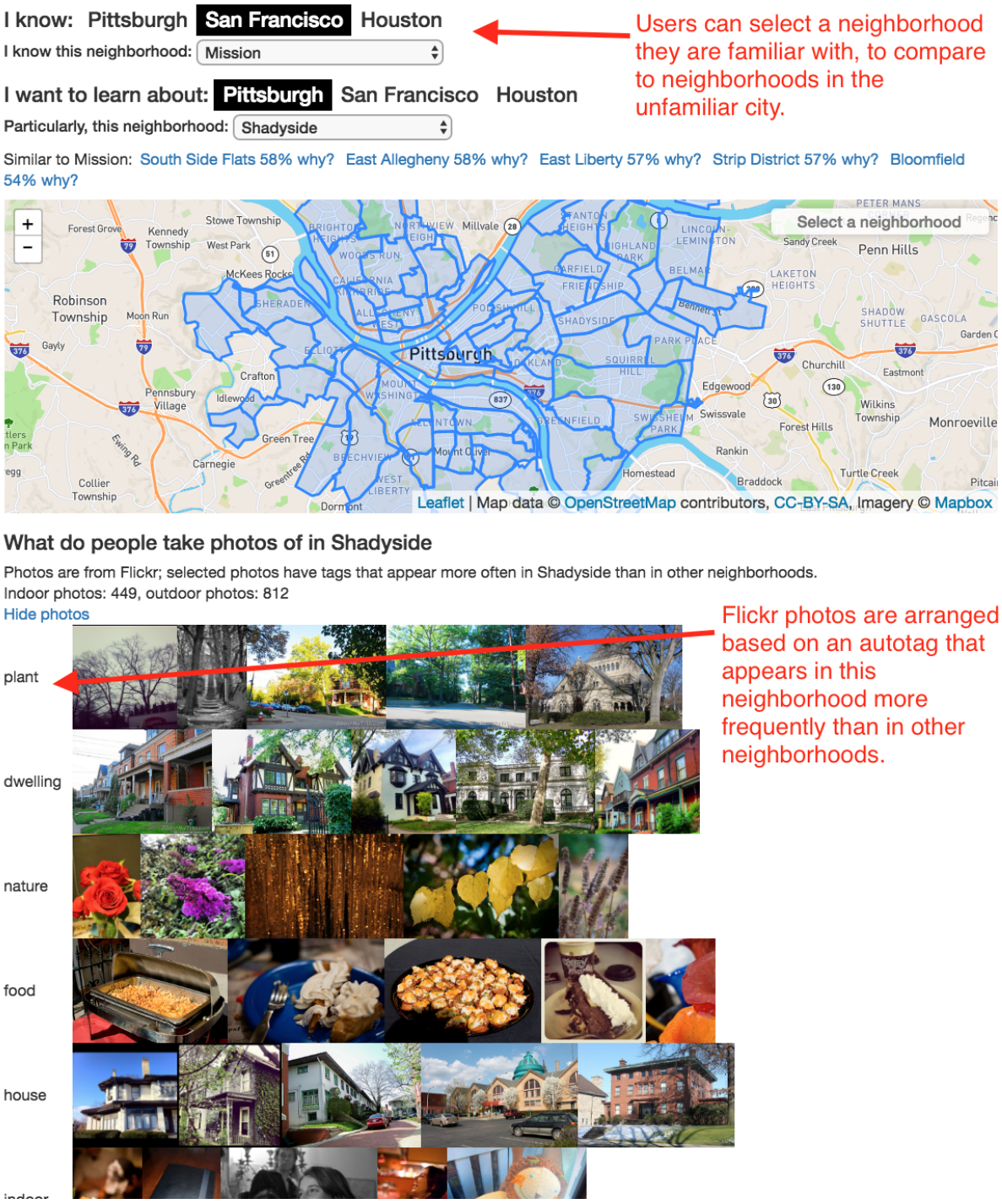


Figure 5.1: Neighborhood Guides prototype v1.0 screenshot, part 1 of 3

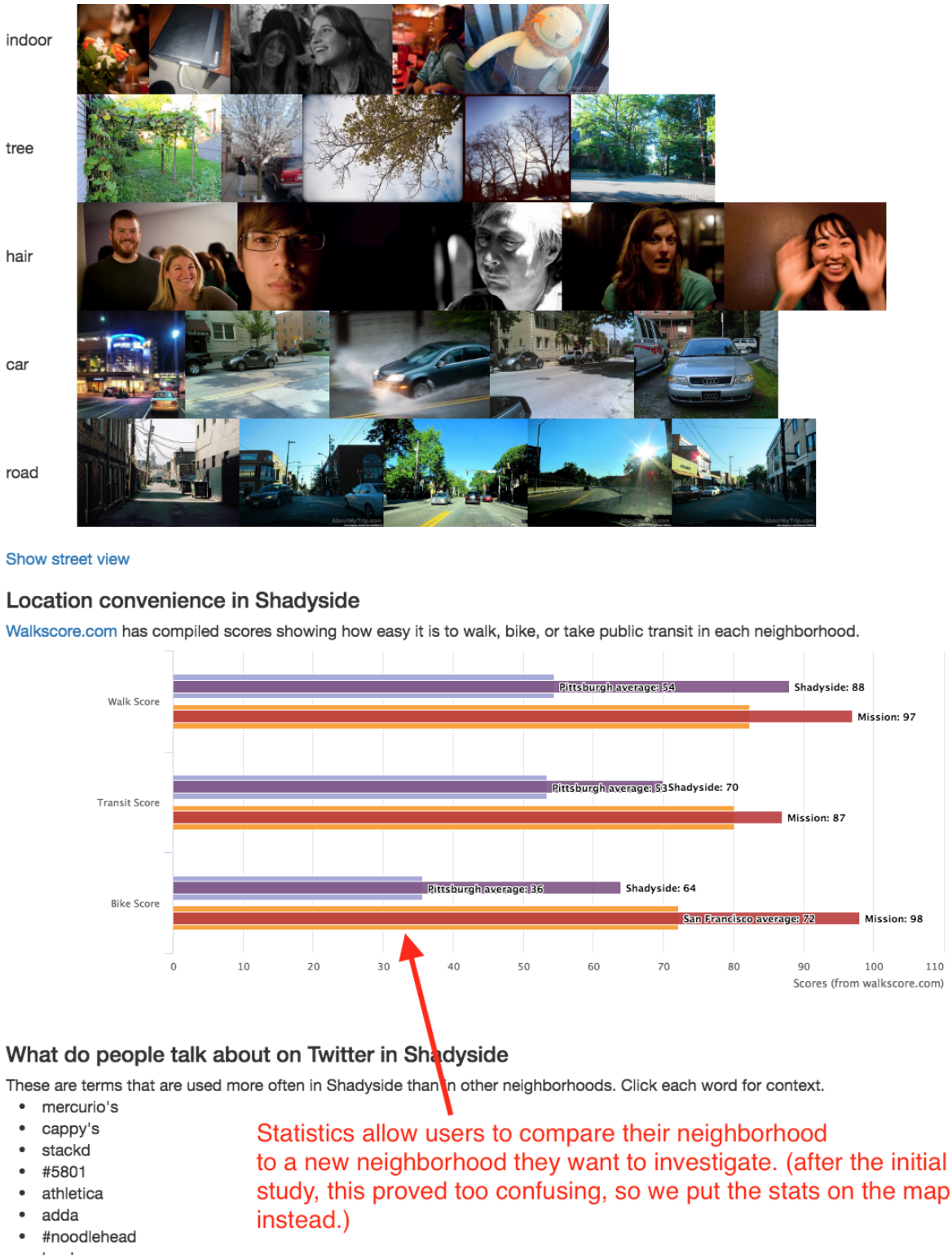


Figure 5.2: Neighborhood Guides prototype v1.0 screenshot, part 2 of 3

What do people talk about on Twitter in Shadyside

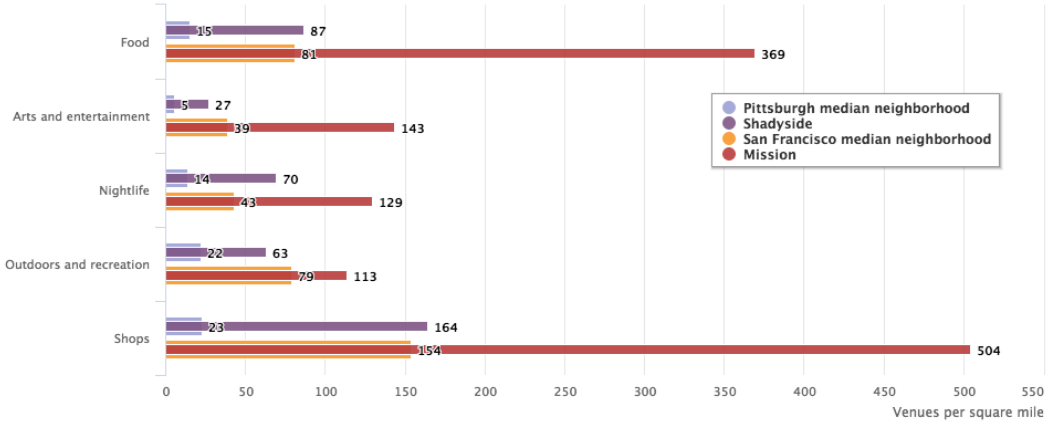
These are terms that are used more often in Shadyside than in other neighborhoods. Click each word for context.

- mercurio's
- cappy's
- stackd
- #5801
- athletica
- adda
- #noodlehead
- kards
- pallantia
- chophouse

Users could click on these terms to get more context about each term.

How lively is Shadyside

Venues per square mile (via Foursquare):



Crime in Shadyside

Part 1 and Part 2 crimes are defined by the FBI. More info
For some cities, only Part 1 crime data is available.

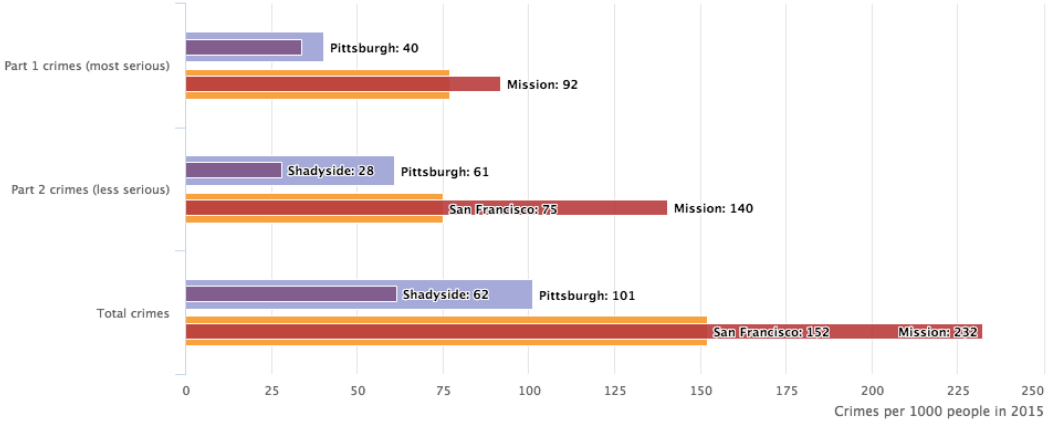


Figure 5.3: Neighborhood Guides prototype v1.0 screenshot, part 3 of 3

5.2.2 CONVENIENCE: WALKSCORES

The company Walkscore[6] has built a Walkability index showing how easy it is to get to local businesses and amenities within walking distance. Their exact algorithm is proprietary, but according to their website, the metric is calculated by counting the number of amenities in different categories within a 5-30 minute radius, where amenities within 5 minutes give the maximum score and amenities outside 30 minutes do not add to an area's score at all. A Walk Score of 90-100 means "Daily errands do not require a car," while 0-24 means "Almost all errands require a car."[7] They similarly compute a Transit Score and Bike Score based on how easy it is to get around using public transit and a bicycle.

In addition to providing a site where users can search for the Walk Score of any particular address, Walkscore also publishes average Walk Scores of entire neighborhoods. We used this data for our prototype. If the neighborhoods that Walkscore uses did not match our official neighborhood boundaries, we adjusted them manually to bring the neighborhood lists into agreement. (For example, sometimes we combined two Walkscore neighborhoods, and averaged their scores, to match one official neighborhood.) Walk Score is a rough metric, but previous studies show that Walk Score explains as much variance for shopping as any other metric, and is comparable with other walkability indices by other measures [58].

5.2.3 AESTHETIC APPEAL: PHOTOS FROM FLICKR

Aesthetic appeal is a rather subjective measure. Because different people have different views on what a "beautiful" place looks like, we cannot distill it into a statistic as we distilled crime and walkability. Therefore, we instead focused on providing users with a representative set of photos of the neighborhood, so they could decide if they liked how it looked.

To do this, we used data from the YFCC100M dataset [89], as it contained 49 million geotagged photos. It also contains "autotags" for each photo: tags denoting what is in the photo, added automatically by a computer-vision-based system. These tags, like "dog" or "building", let us not only pick out photos in a given neighborhood, but also photos of a given subject in a given neighborhood. This, then, let us pick out autotags that were unusually common in a given neighborhood. We did so using the following algorithm.

Let $Score(t,n)$ be the score for autotag t in neighborhood n. Let $C_{t,n}$ be the count of photos that have autotag t in neighborhood n. Similarly, C_t is the count of photos that have autotag t, C_n is the count of photos that are in neighborhood n, and C is the count of photos in the entire city. Then we define:

$$Score(t,n) = \frac{C_{t,n}}{C_n} - \frac{C_n}{C} \tag{5.1}$$

In other words, the score for an autotag t in neighborhood n is the percent of photos that have t in that neighborhood, minus the percent of photos that have t in the whole city. We did this to

[6] https://www.walkscore.com

[7] https://www.walkscore.com/methodology.shtml

penalize tags like "outdoor" and "indoor" that are common in all photos, while still retaining the tags that are most common in each neighborhood. We used this instead of a TF-IDF based algorithm because the autotags come from a closed vocabulary of about 1700 words, so the long tail of natural language does not apply. (Preliminary experiments with TF-IDF variants showed that this was the case; TF-IDF tended to privilege common tags like "outdoor" because the number of "outdoor" photos was so high, even when divided by the number of other neighborhoods that have "outdoor" photos.)

For each neighborhood, we choose the ten autotags with the highest scores, and randomly select five pictures from that neighborhood that have that autotag. We display the photos and the autotag as shown in Figures 5.1 and 5.2.

5.2.4 LIVELINESS: FOURSQUARE VENUES

Our participants described wanting to be in a lively place, where things were happening and there were businesses and people around. While we cannot measure this exactly, we used the presence of Foursquare venues as a reasonable proxy. Foursquare venues are usually businesses such as restaurants, bars, cafes, and shops (despite some occasional creative uses [53]), so a place with more venues is likely to be more lively.

We used the Foursquare Venues API[8] to find all the venues in each city, then divided them into neighborhoods. We used the top-level venue categories to further divide venues into "Food", "Arts & Entertainment", "Nightlife Spot", "Outdoors & Recreation", and "Shop & Service" venues. The other top-level categories ("Travel & Transport", "Residence", "College & University", "Professional & Other Places", "Event") were discarded, either because they would not add to liveliness of the area or because there were very few venues of that type.

5.2.5 LOCALNESS: TWITTER COMMON WORDS

This section of the website is, by necessity, the most subjective. If participants want to "live like locals," we wanted to give them a way to see what locals talk, do, and think about. Twitter is not a perfect sample of this, but it is one sample, and it's free and publicly available, so we drew our data for this section from it.

We had been storing coordinate-geotagged tweets from Pittsburgh since January 2014, San Francisco since June 2014, and Houston and Chicago since November 2014. Details of our data collection are shown in Table 5.2. At the time we built this tool, this meant we had between 2 and 3 years of all of the coordinate-geotagged tweets (tweets with a latitude-longitude point, as in Chapter 3). We began by removing all tweets from the accounts of known spammers, as job-posting spammers had become a significant portion of our data (see Chapter 3 for more details). Specifically, we removed any account that contained "job", "career", "tmj", "join", "workat", "recruit", or "soliant", as those seemed the worst offenders. We split each tweet into tokens using a python port[9] of the Twokenize tokenizer [62]. We then removed links, @-references, tokens that are all punctuation, and stopwords using the list provided in NLTK [9]. Finally, we combined all words that remained in each neighborhood into an unordered bag of words.

[8] https://developer.foursquare.com/docs/venues/search

[9] https://github.com/myleott/ark-twokenize-py

City	Latitude bounds	Longitude bounds	Start date	Approx. # tweets
Pittsburgh	40.241, 40.641	-80.2, -79.8	Jan 2014	4.6 million
San Francisco	37.565, 37.865	-122.595, -122.295	Jun 2014	6.2 million
Houston	29.550, 29.958	-95.592, -95.138	Nov 2014	6.5 million
Chicago	41.637, 42.037	-87.985, -87.519	Nov 2014	7.4 million

Table 5.2: Details of our Twitter data collection in the cities we used for our prototype. For each city, we collected coordinate-geotagged tweets in a rectangle described by these latitude-longitude coordinates.

Using these bags of words, we calculated the score for each word in each neighborhood using TF-IDF [78]. Specifically, we define $TFIDF(w,n)$ for word w in neighborhood n as follows: Let $C_{w,n}$ be the number of times w appears in neighborhood n, and let N_w be the number of neighborhoods w appears in. Then:

$$TFIDF(w,n) = \frac{\log(C_{w,n})}{N_w} \tag{5.2}$$

For each neighborhood, we identify the 10 words with the highest score, and present each word along with 10 random tweets containing that word for context. An example of how these are displayed is shown in Figure 5.4.

5.2.6 Clarification on splitting data into neighborhoods

For all of these dimensions besides Convenience, we needed to split many points into their neighborhoods. In order to do this, we used the Point-in-polygon implementation in the Python Shapely library[10]. However, this algorithm is slow to use on millions of points and dozens of polygons, so we precomputed a grid of 0.001 degrees latitude/longitude across each city and computed which neighborhood each point here is in. Then, to find which neighborhood a given crime/photo/tweet/venue is in, we rounded it to the nearest grid point and returned that answer. This means we may have some points on a border between neighborhoods, up to 0.0005 degrees away, categorized as the wrong neighborhood. Because 0.0005 degrees is about half a small city block, we decided that was an acceptable inaccuracy.

The code used to generate this grid and find the neighborhood for points is available on Github[11].

5.3 Navigation: Neighborhood Comparison

The previous section described how we generate information about a given neighborhood. However, imagine a user staring at the site for the first time. How would they even decide which neighborhood to look at? Following past research [51] and popular press [74], we decided to

[10] http://toblerity.org/shapely/manual.html
[11] https://github.com/dantasse/pointmap

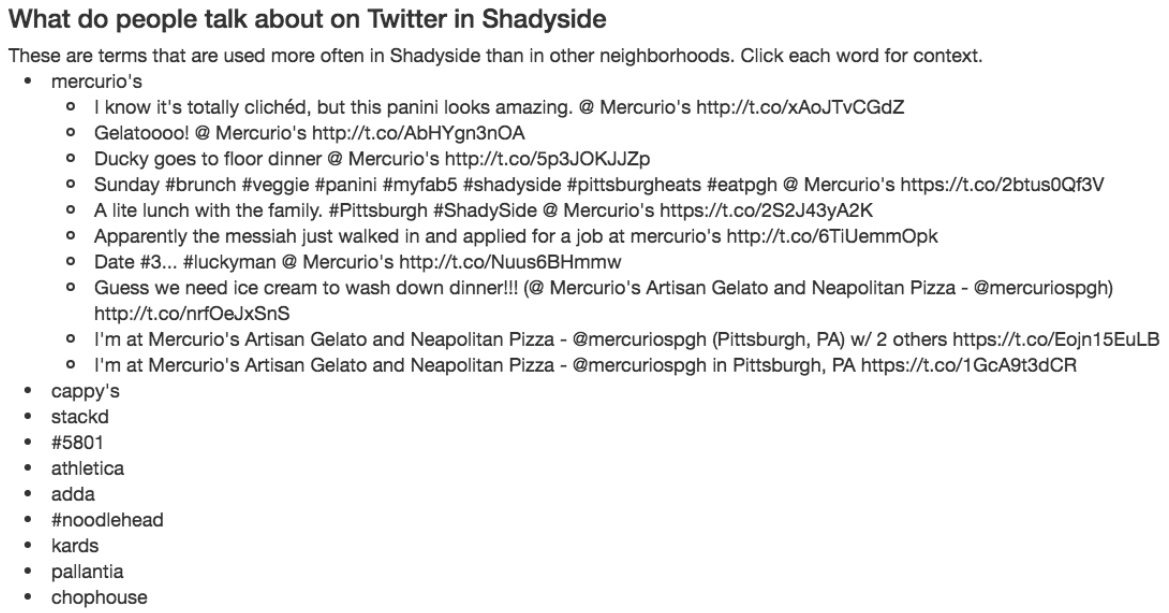


Figure 5.4: The top 10 highest-scoring words in a neighborhood. Users can click on any word to expand or collapse the 10 "context" tweets.

do this by a "similar neighborhood" system. Using this, people could pick a neighborhood in a city that they know, and see recommendations of cities in other neighborhoods. We also chose this path in order to diversify people's choices: perhaps one person might want a neighborhood where they can go out and party, while another person might want a quieter or more family-friendly neighborhood.

To do this, we defined "dissimilarity" measures on each of the five dimensions of data we gathered. We do not claim these measures to be anything other than somewhat arbitrary initial attempts. They seem sensible and they work reasonably well. In addition, there is no meaningful ground truth; no canonical source of "The Williamsburg of Pittsburgh[12]," so any attempt to optimize these would be necessarily imprecise.

We defined dissimilarity of crime as one minus the ratio of crime rates in each neighborhood. If they have similar crime rates, their dissimilarity will be 1; if their crime rates are very different, their dissimilarity will be 0. Formally, let $Cr(n)$ be the total number of crimes in a neighborhood. Then:

$$D_{crime}(n_1, n_2) = 1 - \min\left(\frac{Cr(n_1)}{Cr(n_2)}, \frac{Cr(n_2)}{Cr(n_1)}\right) \tag{5.3}$$

We defined dissimilarity of Walk Scores as the Euclidean distance between the vector of Walk

[12]Though obviously it's Lawrenceville.

Scores. Formally, if $WS(n)$ is the Walk Score of neighborhood n, $TS(n)$ is the Transit Score, and $BS(n)$ is the Bike Score, then:

$$D_{walkscore}(n_1,n_2) = \sqrt{(WS(n_1)-WS(n_2))^2+(TS(n_1)-TS(n_2))^2+(BS(n_1)-BS(n_2))^2} \quad (5.4)$$

Similarly, dissimilarity of Foursquare venues is the Euclidean distance between the vector of Foursquare venues. So if $V_{food}(n)$ is the number of food venues in neighborhood n, then:

$$\frac{D_{venues}(n_1,n_2) = \sqrt{\left(V_{food}(n_1)-V_{food}(n_2)\right)^2+\left(V_{nightlife}(n_1)-V_{nightlife(n_2)}\right)^2+\ldots}}{\left(V_{arts}(n_1)-V_{arts(n_2)}\right)^2+\left(V_{shops}(n_1)-V_{shops(n_2)}\right)^2+\left(V_{outdoors}(n_1)-V_{outdoors(n_2)}\right)^2} \quad (5.5)$$

Dissimilarity of Flickr photos was defined as the sum of absolute differences of the occurrence of each autotag, divided by the sum of both vectors. Defining $C(t,n)$ as the number of times tag t appears in photos in neighborhood n:

$$D_{photos}(n_1,n_2) = \frac{\sum_{t\in tags}|C(t,n_1)-C(t,n_2)|}{\sum_{t\in tags}C(t,n_1)+\sum_{t\in tags}C(t,n_2)} \quad (5.6)$$

We defined dissimilarity of tweets (D_{tweets}) as one minus the cosine similarity between the corpora of all tweets in each neighborhood, as implemented in the Gensim python library [75].

Finally, we defined overall dissimilarity between two neighborhoods as the average of these 5 similarities:

$$D(n_1,n_2) = \frac{D_{crime}(n_1,n_2)+D_{walkscore}(n_1,n_2)+D_{venues}(n_1,n_2)+D_{photos}(n_1,n_2)+D_{tweets}(n_1,n_2)}{5} \quad (5.7)$$

The neighborhoods with the lowest dissimilarities to the neighborhood chosen by the user were shown as the "Similar Neighborhoods," and users could click "why?" to see each of the 5 ratings that went into that average. (See the top of Figure 5.1.) We also used the neighborhood that the user chose in order to show comparable statistics for the Safety, Convenience, and Liveliness sections (See charts in Figure 5.2 and 5.3). Our goal was to show the user not only the raw numbers of crime rates, walk scores, and venue counts, but also to show the user some stats to compare to: statistics for the city as a whole, statistics for the neighborhood they know well, and statistics for the whole city they know well.

5.4 Initial Reactions

Our ultimate evaluation would help us find answers to our latter two research questions: can social media help creative tourists, and what does social media tell us about our cities? However, we first conducted an initial user test in order to catch any obvious mistakes and iteratively improve our prototype. This first evaluation was conducted with 10 travelers in public places in

	Age	Occupation
C1	25-29	Retail
C2	55-59	Test Driver
C3	55-59	Paralegal
C4	30-34	Community Organizer
C5	25-29	Designer
C6	20-24	Finance
C7	30-34	Research Scientist
C8	25-29	PhD Student
C9	60-64	Unemployed
C10	30-34	Research Engineer

Table 5.3: Participants in prototype initial evaluation.

or near San Francisco. Participants were paid $15 for a session between 30-60 minutes in which I would talk with them briefly about their travel experience, describe the site, and ask them to use it to plan a hypothetical trip. After that, I would ask them, among other questions, to order the five parts of the site in usefulness. Participants are described in Table 5.3. This study was approved by the Carnegie Mellon Institutional Review Board.

5.4.1 PHOTOS ARE THE MOST USEFUL; TWEETS ARE THE LEAST

All users ranked the parts of the site, except for participant 9, who declined. We scored rankings by giving 1 point to their most useful section, up to 5 points for their least useful section. If they found two equivalently useful, we gave them both the average of the two scores. (That is, if Photos and Venues were tied for first place, they would both get 1.5 points.) We then averaged across all 9 participants, ending up with the ranking in Table 5.4.

Photos were almost universally preferred as the most useful information source. Participants C5 and C6 described that the photos gave them the best flavor or feel; C7 described appreciating them because they were unfiltered. As she said, “it’s easy to browse through and get a feel without having to read anything and get other people’s opinions.” Similarly, they have noticeable practical uses: in the hypothetical trip planning, C1 and C3 were both saved from neighborhoods that were more suburban and boring than they thought.

On the other hand, tweets were seen to be the least useful. C10 described how the words are useless without context, and sometimes the context provided failed to even give him the necessary context. Crime data was seen as almost as useless, but its distribution is bimodal. In interviews in Chapter 4, some described crime data as being the most important to them, while others did not care at all, and the same bimodal distribution held here. C6 thought crime data was the most

Section	Rank	1s	2s	3s	4s	5s
Photos	2.16	5	1	0.5	1.5	1
Walk Scores	2.33	2	3	3.33	0.33	0.33
Venues	3.00	1	3	2.33	0.33	2.33
Crime	3.67	1	0	2.33	3.33	2.33
Tweets	3.83	0	2	0.5	3.5	3

Table 5.4: Participants' average usefulness rank of each section in the initial evaluation. Lower is better (e.g. Rank=1 would mean that everyone ranked this section as the #1 most useful section). The 1s through 5s column reflect how many people picked this data source as their 1st, 2nd, etc. rank. When participants said ranks were equal, we split their rankings among the categories.

important, while many others found it 4th or 5th most important.

5.4.2 Neighborhood comparison may be too complicated and situational

As described above, we used neighborhood comparison to help users navigate through the list of neighborhoods. At the same time, I tried to ask participants if they would rather have simple "Top 5 neighborhoods" lists, such as "Top 5 Arts Neighborhoods" or "Top 5 Nightlife Neighborhoods" instead of the similar neighborhoods. Four participants (C1, C4, C5, and C6) said yes, while two (C8 and C10) said maybe. Nobody preferred the similar neighborhoods.

Neighborhood comparison may have failed for three reasons. First, confusion: as C10 said, "Similar to Mission" is tricky to interpret because there are so many things that could mean. C6 brought up the related point of trust: as soon as the algorithm "failed" (gave her a comparison that she knew to be wrong) once, she would stop trusting it. Finally, C1 brought up another interesting objection: he travels to find something different, not something the same as home.

Regardless of reason, the comparison between home neighborhoods and neighborhoods in the target city proved less useful than we hoped. Given its lack of utility, we decided it would be better to put statistics on a colored chloropleth map, rather than showing charts for statistics separately.

5.4.3 Details can color users' experiences

Photos serve users as tiny windows into other people's lives, helping them to understand a deep (though narrow) window. Sometimes users would fixate on this window and generalize that to the whole area. For example, C1 noticed a tweet about "elderberry syrup" in Northwest Pittsburgh and thought that that would be an interesting area to explore. Unfortunately, this was not from a business that sold elderberry syrup; it was just someone at their home. If he went there looking for elderberry syrup, he would probably be disappointed. Similarly, C10 saw some photos of a hockey game in downtown Pittsburgh and fixated on hockey, thinking about that as the main thing to do in Pittsburgh. Downtown Pittsburgh has hockey games, but there is certainly more to

the area! Even names can give people this kind of tunnel vision, as when C8 saw a neighborhood called "Museum Park" in Houston and decided it must be stuffy and boring without looking through it thoroughly. This emphasized for us the importance of a diverse set of data so as not to imply that something is more prevalent than it is.

5.5 CONCLUSION

In this chapter, we describe our first Neighborhood Guides prototype and an initial user study based on it. We gained some valuable preliminary insights that will help us design the second iteration. In addition, we learned that the primary data source that will help travelers learn about neighborhoods is photos, which guided our next study.

6 WHICH PHOTOS BEST REPRESENT EACH NEIGHBORHOOD

When travelers are viewing data from cities remotely, they have many data sources to choose from. Between tweets, venues, statistics, photos, and other data sources, it is intuitive that photos would be the most helpful, because they are the richest. In addition, recent advances in computer vision (like those described in [89]) are increasingly making it easy to sort, process, and otherwise use these photos. In Chapter 5, initial reactions suggested that photos would be a valuable part of any tool that used social media to help travelers, as well.

However, which photos should we use? In the YFCC100M dataset alone, there were roughly 900,000 photos from San Francisco, which translates to roughly 20,000 per neighborhood. Furthermore, there are more photo sources beyond Flickr. Previous work has investigated ways to summarize sets of geotagged data [4, 37, 42, 43], but are these methods ideal for summarizing a place for travelers? Finally, is social media data the best data to describe a neighborhood, or would users be better off with something like Google Street View photos that more accurately shows the city from the street?

To answer these questions, we devised a Mechanical Turk experiment in which users could tell us which data sources best represented their neighborhoods. We tested six data sources: three from Flickr, two from Street View, and one from Instagram. We gave them two pairs of photos and asked participants to compare which photo set better represents their neighborhood, then aggregated to determine the overall best photo sets.

6.1 STUDY METHODS

We recruited participants from Amazon Mechanical Turk who have lived in, or were otherwise very familiar with, Pittsburgh, San Francisco, Chicago, Houston, or Austin, as those were the cities we were able to gather data from. After the consent form, we asked them which neighborhood they had lived in or otherwise knew well. We then showed them the six photo sets described in Section 6.1.1, two at a time, and asked them which one they thought better represented their neighborhood. A screenshot is shown in Figure 6.1.

The survey took up to 5 minutes, and participants were paid $1.00. We recruited 200 participants who had at least 1000 HITs accepted on Mechanical Turk and at least a 98% previous acceptance rate. After inspecting their work, we found no reason to reject any workers.

6.1.1 PHOTO SETS

For each neighborhood, we collected the following photo sets:

1. STREET VIEW RANDOM: A random set of 10 photos from Google Street View in the neighborhood. We collected these because a few participants in the initial evaluation (in

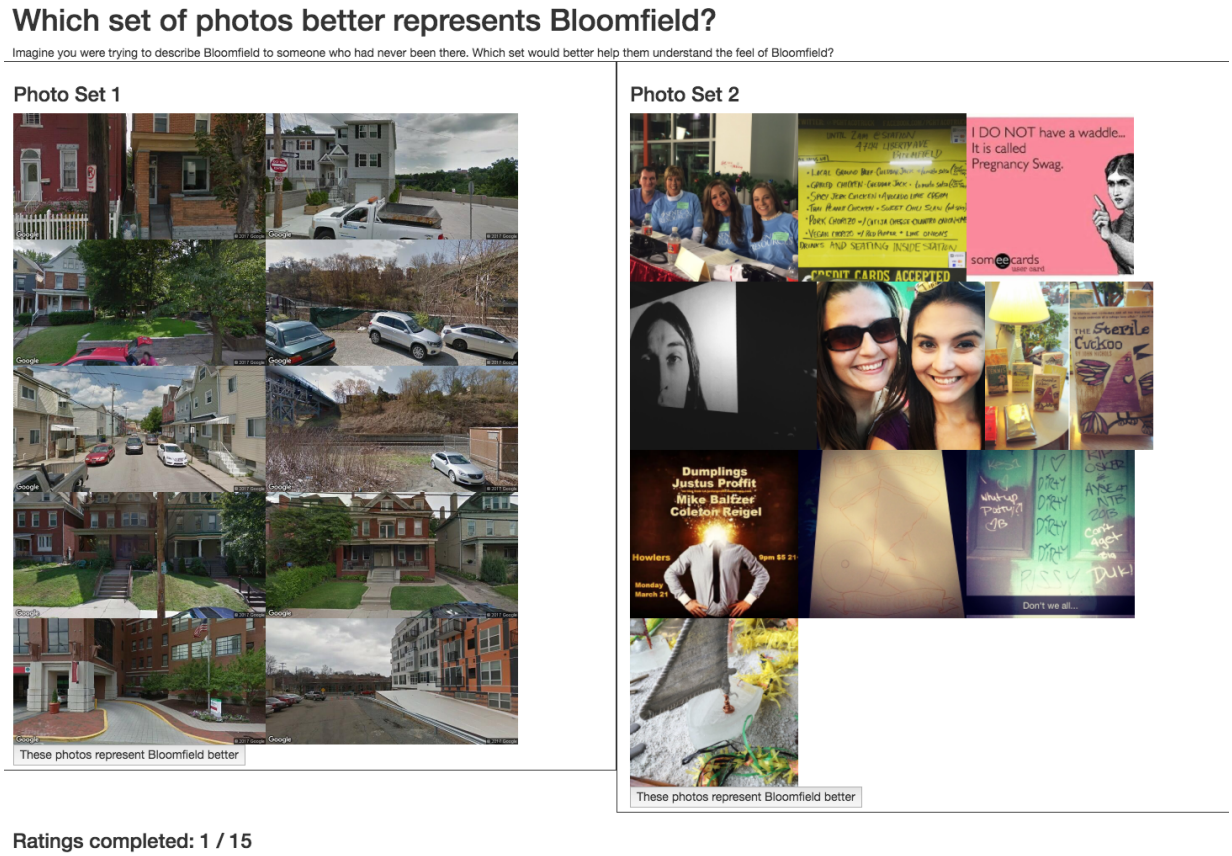


Figure 6.1: A screenshot of our study interface in this chapter's study. Users would see 15 of these pairs of photo sets: all combinations of the six photo sets described in Section 6.1.1.

Chapter 5) suggested that Street View might be a good way to view a neighborhood.

2. STREET VIEW VENUES: A set of photos from Google Street View near Foursquare venues. To do this, we randomly picked 10 Foursquare venues from the neighborhood, then picked the nearest Street View photo to each. We picked these because participants suggested that knowing the commercial areas is more useful than just residential areas.

3. FLICKR RANDOM: A set of 10 random Flickr photos from the YFCC100M data set in the neighborhood. These comprise a baseline social media photo set.

4. FLICKR ONE-PER-USER: A set of 10 random Flickr photos from the YFCC100M data set, with the constraint that no more than one could come from each photographer. We selected these because a few participants in our initial evaluation noticed situations where one photographer's photos made up most of the photos in a data set, so they did not get the full sense of the diversity of viewpoints in the neighborhood.

5. FLICKR JAFFE: A set of 10 Flickr photos, chosen according to the method in [37]. This method involved using the Hungarian clustering algorithm [26] and then ranking each cluster based on tag distinguishability, photographer distinguishability, photo density, and image qualities.

6. INSTAGRAM: A set of 10 Instagram photos. Because Instagram has no public photos API, we randomly selected these from the Instagram photos that were cross-posted to Twitter. We included these to see if there were significant differences between photos shared on different platforms.

6.2 RESULTS

Of the 200 participants, 61 of them selected a neighborhood in Chicago, 51 in San Francisco, 42 in Pittsburgh, 23 in Houston, and 23 in Austin. We used the Bradley-Terry Model [10] to convert pairwise comparisons into a complete ordering among the six photo sets. The model can be interpreted as follows. If the score for element i is p_i and the score for j is p_j, then the probability that i will win against j in a pairwise competition, denoted $P(i > j)$, is

$$P(i > j) = \frac{p_i}{p_i + p_j} \tag{6.1}$$

6.2.1 STREET VIEW PHOTOS WERE MOST REPRESENTATIVE

Our main finding was that Street View photos were most often found to be the most representative. See Figure 6.1 for details. Each of the Street View photo sets handily defeated all of the social media photo sets. In addition, we were surprised to see random Flickr photos performing better than the photos selected according to [37].

6.2.2 SOCIAL MEDIA PHOTOS DID BETTER IN ICONIC NEIGHBORHOODS

In order to understand this counterintuitive finding, we investigated each of the neighborhoods that had at least 4 participants choose it (see Table 6.2). In five of them (Chinatown, Golden

Photo Set	Bradley-Terry Score	Win Percentage
STREET VIEW VENUES	0.233	61%
STREET VIEW RANDOM	0.219	59%
FLICKR RANDOM	0.159	50%
FLICKR ONE-PER-USER	0.148	47%
FLICKR JAFFE	0.124	42%
INSTAGRAM	0.116	40%

Table 6.1: Bradley-Terry Model Scores and win percentages for each photo set. Higher Bradley-Terry parameters indicate that they won more comparisons.

Gate Park, Haight-Ashbury, Wrigleyville, and Downtown Houston), social media data sets did outperform Street View. We noticed that these were mostly rather iconic neighborhoods. Chinatown is known for its unique decorations and excellent restaurants; Golden Gate Park is the largest park in San Francisco, with many attractions; Haight-Ashbury has a storied history with hippies and the 1967 Summer of Love; Wrigleyville is the home of Wrigley Field, the legendary Chicago Cubs ballpark. The photo sets that users selected tended to emphasize the iconic nature of these neighborhoods (e.g. photos of Chinese food in Chinatown; photos of the stadium in Wrigleyville), while Street View photos showed humdrum streets. See Figures 6.2 and 6.3 for examples of why Chinatown was better reflected in its Instagram photos.

6.2.3 SOMEWHAT MORE BUILDINGS IN HIGHER-SCORING FLICKR DATA

Because the Flickr photos from the YFCC100M dataset have autotags attached that reflect their content, we can analyze this content. We selected the "winner" Flickr photo sets, defined as those that won at least 75% of their comparisons, and compared their autotags to the autotags of all photo sets, looking for tags that appear much more or less often in the "winner" photo sets. Results are shown in Table 6.3. It appears that tags relating to the built environment were more prevalent in the "winner" photo sets, while tags about nature were less prevalent.

6.3 DISCUSSION

In this study, we found that these users think Street View photos represent their neighborhood better than social media photos. However, there are a few higher-level points we want to suggest.

6.3.1 VISITOR'S EYE VS. LOCAL'S EYE

There may be a disparity that we didn't anticipate between users who have lived in a place and users who know a place well. The 7 participants who selected "Golden Gate Park," which has no housing, suggests that a substantial number of people selected neighborhoods that they knew well, rather than neighborhoods that they lived in. In addition, Wrigleyville and Chinatown are not large neighborhoods, so we would be surprised if many people lived there.

Neighborhood	City	Participants	Most popular photo set	Is iconic?
Chinatown	San Francisco	8	INSTAGRAM	Yes
Golden Gate Park	San Francisco	7	FLICKR ONE-PER-USER	Yes
Haight-Ashbury	San Francisco	7	FLICKR RANDOM	Yes
Wrigleyville	Chicago	5	INSTAGRAM	Yes
North Austin	Austin	5	STREET VIEW VENUES	No
Allegheny Center	Pittsburgh	5	STREET VIEW RANDOM	No
Downtown	Houston	4	FLICKR RANDOM	No
Lake View	Chicago	4	STREET VIEW VENUES	No
Pacific Heights	San Francisco	4	STREET VIEW RANDOM	No

Table 6.2: Most popular photo sets in neighborhoods with at least 4 raters.

Autotag	Frequency difference
`architecture`	5.7%
`building`	4.3%
`building complex`	2.1%
`skyscraper`	1.7%
`road`	1.5%
`indoor`	–4.1%
`outdoor`	–3.8%
`nature`	–2.5%
`plant`	–2.0%
`animal`	–1.5%

Table 6.3: Flickr autotags that were most and least prevalent in the most successful photo sets. Frequency difference is the percent of photos that had this tag in the "winner" photo sets minus the percent of photos that had this tag in all photo sets. Therefore, high frequency difference indicates that this tag appears in very representative photos, while low frequency difference indicates that this tag does not appear often in very representative photos.

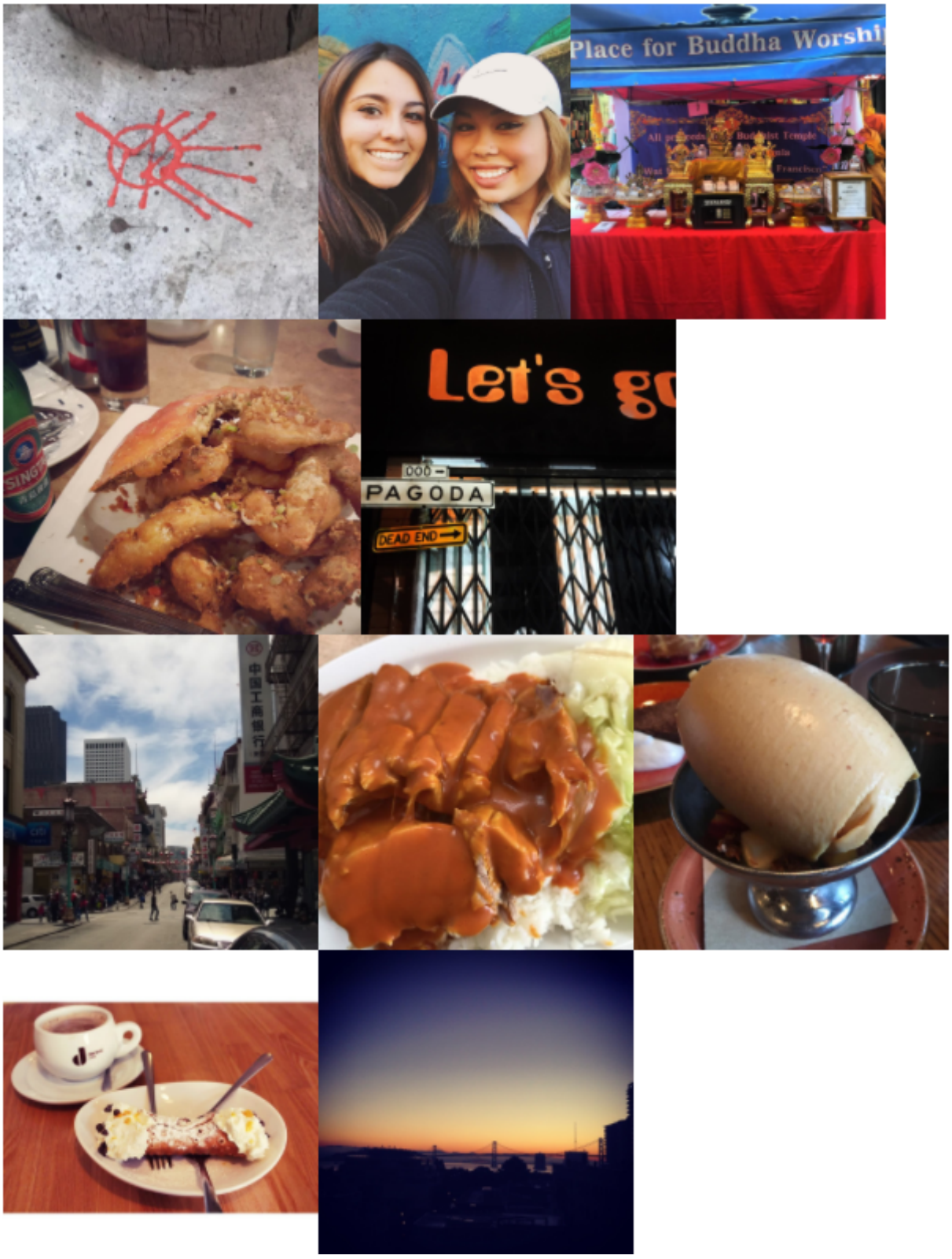


Figure 6.2: The photos from Instagram (INSTAGRAM) in Chinatown, San Francisco. Notice how they reflect the neighborhood's interesting character, especially the wide variety of food available.

Figure 6.3: The photos from Google Street View (STREET VIEW RANDOM) in Chinatown, San Francisco. Most of these reflect boring-looking buildings, because the most interesting parts of Chinatown are inside these buildings.

If we assume that these users were visitors, it may suggest an explanation of why their results were different than the majority. When they went to Wrigleyville, they may have seen the baseball stadium, so people's photos within the stadium seem "more representative" than photos of the street outside. People who live in Wrigleyville may think the opposite.

This also highlights the distinction between photos that are representative and photos that are useful. When travelers are going to a new city, they may not want to have a perfect representation of that city. This brings us back to the "ideal everyday" idea from Chapter 4.5.4. These travelers want to understand a slice of everyday life, but not the boring parts of everyday life. They don't particularly want to go to an office hour 8 hours, or renew their driver's license at the DMV. They want to see what part of everyday life is there that is not everywhere else.

Locals, on the other hand, may have enough experience there to see things more realistically. They may see the photos from the street as the "real" city, while the attractions are only the "tourist" city.

6.3.2 RANDOM IS OK

One side finding that intrigued us was the fact that the model from Jaffe et al [37] did not outperform random Flickr photos, and the STREET VIEW VENUES model did not strongly outperform random Street View photos. This may be another side effect of asking about what is "most representative" instead of what photos are most individually useful. It may also be the case that it matters less what the content of each individual photo is and more that there simply are some photos.

While this study did not give us the result we expected, it gave us an interesting avenue to explore. Perhaps we should include street view photos in our neighborhood guides, or perhaps we worded this study poorly and ended up investigating something other than we wanted. The findings that more iconic neighborhoods being reflected better in social media reflects the latter, but our final study will hopefully elucidate this point.

6.3.3 THE BUILT ENVIRONMENT IS MORE REPRESENTATIVE THAN NATURE

Our findings in Table 6.3 suggest that tags like `Building` and `Architecture` appeared more often in photo sets that were selected as winners, while tags like `Nature` and `Plant` appeared less often. This makes sense when one considers that built environment has more variance; nature probably looks similar throughout a city, while buildings can vary widely.

Another interesting finding is that both `Outdoor` and `Indoor` tags, two of the most common tags throughout the YFCC100M data set, are found less often in "winning" photo sets. Perhaps this reflects that photos are more representative if it's hard to tell whether they are indoors or outdoors. The lack of an `indoor` or `outdoor` tag may also mean that there is some action or movement in the photo that makes the exact location harder to determine. Further study into the system tagging these photos would be helpful to understand what these findings mean.

6.4 CONCLUSION

In this chapter, we describe a study we ran on Mechanical Turk in which we asked people to tell which photo sets best represented their home neighborhood. We found that Street View photos were the most representative in most neighborhoods, though some touristically-relevant neighborhoods were best represented by social media photos. We also found that buildings and architecture are likely to make a photo set more representative, while plants and nature are generally less representative.

7 USER STUDY WITH NEIGHBORHOOD GUIDES

After the initial evaluation of the Neighborhood Guides, and the further study to show which photos are likely to be most helpful, I returned to the Neighborhood Guides application with more insights to guide our design. In this section, I describe first the improved Neighborhood Guides site, then detail the study I ran and the insights learned from it. I found that social media photos show the idealized side of the city that creative tourists want to see, that the best photos show people doing something, that statistics are useful but can be greatly simplified, and that people search for textual "blurbs" to give them a schema to base their understanding of the neighborhood on.

7.1 NEIGHBORHOOD GUIDES 2.0

After the feedback from the initial evaluation in section 5.4, we made some changes to the Neighborhood Guides website. In this section, we will describe those changes and why we made them.

7.1.1 FOCUS ON PHOTOS, HIDE TWEETS AND SIMILAR NEIGHBORHOODS

This change was relatively straightforward, given our results in section 5.4.1. Photos seemed the most immediately useful data source for travelers, and tweets were the least useful, so we reoriented our site to focus more on photos and we hid the tweets. The new site would only include the map and photos. In the user study we did with Neighborhood Guides 2.0, tweets were hidden to start. (We did retain the option to show them via a small control at the bottom of the page, for our use in the user study.)

We showed the Street View photos first, after the results of our study in Chapter 6, and also included random Flickr photos with one per photographer (corresponding to the FLICKR ONE-PER-USER condition in that study). We maintained the original set of photos sorted by autotags, as well as Flickr photos selected according to the method in [37] (corresponding to the FLICKR JAFFE condition), and random Instagram photos (the INSTAGRAM) condition, which started hidden but could be shown during the study.

We hid the Twitter common words, but likewise maintained the control that let us show them during the study. Of course, our results from the initial study in Chapter 5 do not prove that there is no useful way to use geotagged tweets to describe an area. Our sample size was only 10 people, and even if we had a larger sample that said the same thing, that would only prove that our method of selecting tweets did not describe an area well. Regardless, we chose primarily to use photos in order to focus our attention.

7.1.2 COMBINE MAP AND STATISTICS

Participants in our first study had some difficulty understanding statistics like Walk Score, crime rates, and venue densities when they were separated from the map. We had chosen to separate them in order to allow users to compare statistics between their familiar neighborhood and a new one, but we understood this was less useful than allowing them to quickly compare across neighborhoods in the same city. As a result, we hid the charts comparing each dimension to their home dimension, and added a chloropleth layer on top of the map to let people explore many dimensions quickly.

The lack of comparison charts weakened the "similar neighborhoods" feature, so we hid that too. This also helped remove some of the noise from the site, as more people found it confusing than helpful. Similarly to the other features, though, we maintained the ability to show it if it would help our conversation.

7.2 STUDY METHODS

We ran this user study as an extended version of the user study in chapter 5. We would talk with the participants a bit about a recent trip, then ask them to plan a hypothetical trip to one of the cities in our site that they had not yet been to. (Only one participant, D4, had been to all of these cities; we asked him to choose the city he was least familiar with.) I asked 19 of the 21 participants (the other two ran out of time) about their preferences of photo subsets by turning on and off each photo set. If it came up in conversation and if time permitted, I would also show them the "Similar Neighborhoods" feature and ask them whether they would prefer that or a "Top 5 Arts/Nightlife/etc neighborhood" feature.

We recruited participants through social media and via the CMU Center for Behavioral and Decision Research Participant Pool. Participant demographics are shown in Table 7.1.

Interviews lasted up to 60 minutes, and participants were paid $15. Sessions were conducted in the cafe in the Gates-Hillman Center of Carnegie Mellon University. Because this was a public place, we could not record the interviews; instead, we took notes as the study progressed, and condensed them to the 2-7 most interesting findings after each session (or after a block of sessions, if there were a few in a row). These notes were then analyzed using affinity diagramming (as described in [8]) to identify common themes.

7.3 RESULTS

7.3.1 FLICKR PHOTOS DRASTICALLY OUTPERFORM STREET VIEW

When we asked users to rank which photos were the most useful to them, they responded as shown in Table 7.2. Flickr photos with categories were the best ranking, and Street View photos were the worst. This seems to directly contradict our results in Chapter 6, but recall that we asked slightly different questions. In Chapter 6, we asked which photos best represented each neighborhood, and we asked about neighborhoods people lived in; in this study, we asked about which photos are most useful to travelers.

When asked why they ranked their choices the way they did, participants offered a few expla-

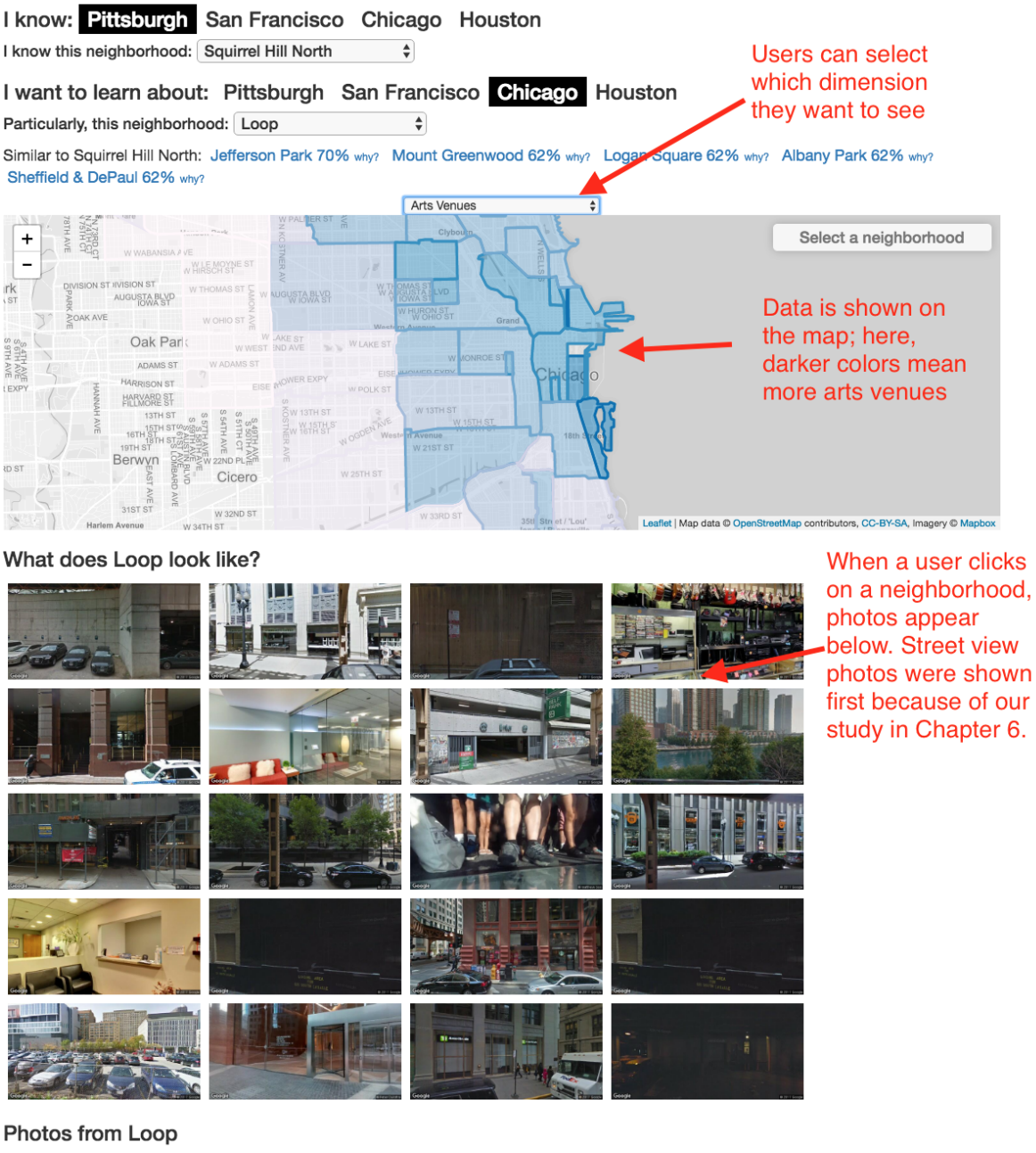


Figure 7.1: A screenshot of Neighborhood Guides 2.0, part 1 of 2. Since version 1, the charts of Walk Scores, crime, and venues have been hidden, replaced by a chloropleth in which users can select different dimensions to display on the map. Street View photos are the first photos on display.

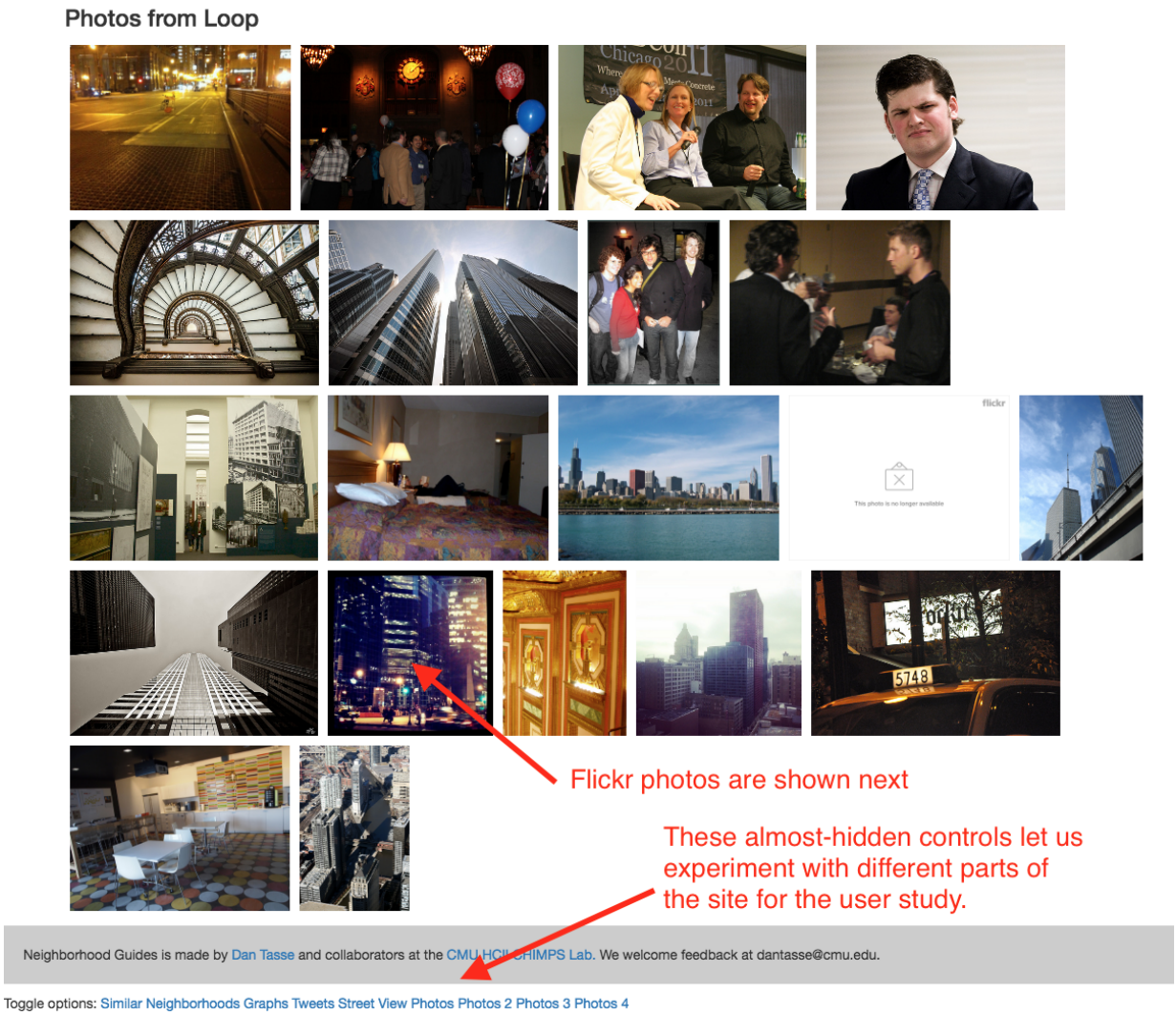


Figure 7.2: A screenshot of Neighborhood Guides 2.0, part 2 of 2. The second set of images is random Flickr images, with a maximum of one image per Flickr user. Notice the control at the bottom to enable us to quickly show or hide different parts of the website.

	Age	Gender	Recruitment method
D1	36-40	F	CBDR
D2	31-35	F	CBDR
D3	25-30	F	CBDR
D4	18-24	M	Social media
D5	41-45	M	CBDR
D6	18-24	F	CBDR
D7	18-24	F	CBDR
D8	51-55	F	CBDR
D9	18-24	F	CBDR
D10	18-24	M	Social media
D11	18-24	M	CBDR
D12	18-24	F	CBDR
D13	18-24	F	CBDR
D14	18-24	M	CBDR
D15	25-30	F	Social media
D16	18-24	F	CBDR
D17	18-24	F	CBDR
D18	56-60	F	CBDR
D19	18-24	M	CBDR
D20	25-30	F	CBDR
D21	18-24	F	Social media

Table 7.1: Participants in this user study. Participants are numbered with 'D' to distinguish them from previous studies.

Photo set	Average rank
FLICKR WITH CATEGORIES	2.33
FLICKR ONE-PER-USER	2.57
INSTAGRAM	3.28
FLICKR JAFFE	3.30
STREET VIEW VENUES	3.53

Table 7.2: Users' average rankings of photo sets. As in Table 6.1, lower is better; "1.00" would indicate that everyone ranked this set their #1 most useful set.

nations. Most of them involved a variant of the idea that they don't actually want to see how a place really looks, they want to see something exciting about the place. D21 described wanting "to capture where people see beauty"; D1 described it as seeing "the best side of each neighborhood." Others described the Street View photos as boring (D7) or "too zoomed-in" (D6). D3, D4, and D12 appreciated the Flickr photos with autotag-based categories, as they help them process all these photos so quickly.

7.3.2 THE IDEAL PHOTO HAS A PERSON, DOING A THING

Many participants talked about wanting to learn not only what a place looks like, but also what people do there. If a photo just has a person's face, that's useless - D7 reported wanting to see people doing things, "not just a picture of a guy." D21 also described how finding activities she could do in the place would be helpful. On the other hand, if a photo just has pictures of the place, that is likewise unhelpful; D3 and D12 described that being a shortcoming of the Street View photos.

On the other hand, people participating in one-time events can be both positive and negative. D1 and D18 enjoyed learning about local events, but D8 and D10 noted that they are only helpful to travelers if the travelers happen to be in town when the event is happening. These events are often causes for taking lots of photos, too, which means that they may become the majority of the photos in a region. Therefore, when selecting photos from a neighborhood, half of the photos might come from something that happens only once a year. This was the case for D8, who found a lot of photos from the Washington Ave Coalition/Memorial Park neighborhood of Houston to be from a running event. She learned little about the neighborhood besides that it had some kind of race once a year.

7.3.3 A "BLURB" GIVES PEOPLE A CONCEPTUAL START

When researching a new neighborhood, participants struggled with trying to make sense of a lot of disparate information. Seeing statistics on a map showed them one side of a neighborhood, but it was impossible to keep them all in their mind; similarly, the photos showed a lot of different stories from different people. They wanted to be able to tell one cohesive narrative about the neighborhood; whatever it was that all these photos and statistics had in common. Of course, any

one narrative would naturally collapse a lot of the neighborhood's natural complexity, but that was fine; it would give them some way to organize all this information in their mind.

D10 described how the Neighborhood Guides site "does a lot at a glance; I kinda want to get the editorialized version." He described how he would read books, local blogs (like "I Heart Reykjavik[1]" when traveling to Iceland), books, or TV shows set in an area before traveling there. He used Wikitravel[2], an online travel guide, to see a factual overview of what's in a place; beyond that, he wanted some description.

Different travelers had different ways to find such a blurb. D15 used the Airbnb neighborhood guide's one-sentence overviews, like "Mexican bakeries, Chinese take out spots, artisanal donut shops, ramen restaurants, and lively bars all near Dolores Park.[3]" D2 and D20 both Googled the neighborhood they were looking at (Hayes Valley, San Francisco, and Montrose, Houston), settling a local newspaper's page[4] and a tourism bureau site[5]. When asked, neither could describe exactly what they were looking for, but they seemed satisfied with those two sites, seemingly because they gave them a fuller picture than a few snapshots.

Lacking even a short blurb, some participants would go so far as to build their own picture, based on a neighborhood's name. This is not necessarily a mistake; for example, D13 described an attraction to the Marina neighborhood in San Francisco because she likes being by the water, and D6 saw Museum Park in Houston and rightly assumed it was where all the museums are. However, it can lead travelers astray, as when D15 saw "Russian Hill" and assumed it must be a heavily Russian area (it's not particularly). More unfortunately, it usually does not give travelers much help, because most neighborhood names are uninformative.

7.3.4 STATISTICS ARE OCCASIONALLY USEFUL

Participants used the map and statistics, but not thoroughly. They would sometimes click through the different maps to get a sense of where anything was, like D2 checking the "All Venues" or D6 trying to avoid "residential" places because there was not as much to do. They would also occasionally use the statistics to avoid high-crime areas, like D20 being intrigued by Midtown but ruling it out after learning of its high crime rate. In these ways, they functioned as simple search tools, directing travelers not to one particular neighborhood but rather to a set of neighborhoods that were at least reasonably dense.

7.4 DISCUSSION

These findings more fully described a lot of compelling concepts from our earlier work, and corrected some misconceptions.

[1] http://www.iheartreykjavik.net/
[2] http://wikitravel.org/en/Main_Page
[3] https://www.airbnb.com/locations/san-francisco/mission-district
[4] http://www.sfgate.com/neighborhoods/sf/hayesvalley/
[5] https://www.visithoustontexas.com/about-houston/neighborhoods/montrose/

7.4.1 PLANNING TRAVEL IS ABOUT GETTING EXCITED AND AVOIDING TRAPS

Clawson and Knetsch, in 1966, described anticipation as the first part of an outdoor recreation experience. As they wrote, "a fisherman may get more enjoyment from tieing his own dry flies through the winter than he will lager get from the actual fishing itself" [12]. This principle rings true in our participants' stories and plans, and can explain many of our findings. A photo showing someone doing an activity would inspire much more anticipation than one simply showing a building or a person; the benefit of the photos is that they can mentally bring people into that activity.

This would also explain some of our other findings. Picture quality and page polish matters (D4 skipped over some sections because they had missing photos; D9 selected as her most useful photo set the one that had the aesthetically best photos). If participants were trying to understand some "true" perception of the city, the photo quality would not matter. But if they are trying to get excited about the area, of course they want to see the most beautiful side of each neighborhood. (Incidentally, this factor may also help explain why Street View photos did better in the study in Chapter 6. They are all the same size, so we could display them in a neat grid, while the social media photos could not be so neatly arranged.)

In addition, the short "blurb" that the participants sought out makes more sense if their goal was to find an exciting part of a city than if they were trying to find an accurate view. No short blurb can explain, for example, all of the history, culture, problems, and triumphs of San Francisco's Mission District, but a blurb can quickly point readers to its "Mexican bakeries, Chinese take out spots, ... and lively bars near Dolores Park."

There is another side to the trip-planning process, however: the necessity of avoiding traps. There are a number of kinds of traps, including unusually high-crime areas, surprisingly spread-out residential neighborhoods, or quintessential tourist traps. This is why D13 talked about being interested in "authenticity;" despite not being able to define it exactly, she knew that "staged" photos from tourist attractions might lead her into a trap. This is also why D19 and D20 changed their plans from their first inclinations (The Tenderloin in San Francisco and Midtown in Houston); the neighborhoods looked interesting, but they realized they might end up in a higher-crime situation than they expected.

Thus, travelers must manage this tension of building excitement while avoiding traps. First person accounts and photos build excitement; by reading individual stories, photos, and blog posts, a traveler can see all the compelling scenarios that might play out when they travel there. However, it is much easier to find what not to do by taking a wider, calmer, birds-eye view like that offered by statistics. In the rest of this chapter, we will explain how these two components can do their jobs optimally.

7.4.2 SOCIAL MEDIA PHOTOS CAN HELP TRAVELERS GET EXCITED

The role social media photos can play for travelers involves building their anticipation about a place. This is why they were the best choices when we asked "what photos are most useful when you're traveling?", while the Street View photos were ranked higher in Chapter 6 when we asked "which photos best represent your neighborhood?" Our participants liked seeing people doing

things, they liked seeing a diverse array of photos, and they liked seeing the most beautiful and well-shot photos in each neighborhood.

The question of which photos should be shown is still an open question. We did not find evidence to support the algorithm of [37], but we did not find any other photo set that was consistently more effective either. We did, however, uncover a lot of pitfalls. Low-quality, low-resolution, or badly-lit photos should be removed. Photos from the same photographer should be limited, and possibly photos from the same day as well; photos are more useful to travelers when they represent a wide and diverse cross-section of the experiences in the area. Some kind of organization seems useful, where our autotag-based categorization is one possible option.

Finally, there is plenty of room to explore when selecting and displaying photos. Some participants speculated about creative ideas like only selecting the photos with text in them (perhaps to show how signs or publications varied from place to place) or to sort photos by focal length to avoid short-range selfies or long-range skyline shots. We encourage future work in this space to investigate new ideas, as long as they promote a diverse array of photos from the neighborhood.

7.4.3 STATISTICS CAN HELP TRAVELERS AVOID TRAPS

Statistics still play an important role in helping travelers, but it is a niche role. We improved the Neighborhood Guides site between versions 1.0 and 2.0 by removing a lot of the screen real estate dedicated to numbers, but we did not go far enough. Our statistics could be further simplified because they are mostly just proxies for residential density: higher downtown, lower in outlying neighborhoods. Clearly the venue densities are this way; Walk Scores tend to follow this pattern too, because it's easier to walk, bike, or take transit when everything is closer together. Crime per capita is not always higher downtown, but it often is; D5 explicitly mentioned that when he saw high crime in the Loop of Chicago. As he said, he wasn't worried about crime downtown because he knows it happens.

Similarly, our removal of comparisons to a user's home neighborhood mattered only in a few edge cases. D10, for example, described how he might want comparisons to his home city only if he's dealing with a particularly high crime region. Despite knowing that a place is the highest-crime neighborhood around, if it is still safer than places he knew, he would not mind.

Therefore, we could replace all of the chloropleth maps with a simple map that showed residential density. This would answer the most common question people used the map for: "where is everything?" This would easily help them avoid the pitfall of booking a place that happens to be in an inaccessible or boring neighborhood, while not overwhelming them with numbers. Perhaps a crime map could be included too; as in the study in Chapter 5, it's something that most people don't care about, but a few people cannot live without.

More investigation should be done into what traps travelers see, and how they can be avoided, in order to maximize the usefulness of maps and statistics. Regardless, it seems clear that statistics should be in a supporting role, with photos and other personal accounts featured in order to build travelers' excitement.

7.5 CONCLUSION

In this chapter, we described the results of an investigation into travel planning when users actually have social media and other data in front of them. We found that there is a role for social media to play, but it isn't as straightforward as finding the right information to accurately describe the area. Social media should be used to build travelers' excitement and anticipation for the trip they're about to take by reflecting the diversity of life and experience that exists in that area. Statistics and other data sources can be used as well, to help travelers avoid traps. In the next chapter, we will further discuss our findings and suggest future work to be done in this space.

8 DISCUSSION AND FUTURE WORK

In these studies, we set out to answer the following questions:

- What do creative tourists mean by "getting a feel for the city"?
- Can social media help them achieve this goal?
- What can social media tell us about our cities and neighborhoods?

Chapter 4 provides our answer to the first question: they want to maximize along five dimensions of safety, convenience, aesthetic appeal, liveliness, and the ability to live like locals. Chapters 3, 5, 6, and 7 address the second two questions. In this chapter, we will first revisit the first question, then tie together the results from all of these studies to more clearly discuss what we have learned with relation to the second two. Meanwhile, we will then discuss what these findings mean for future researchers and developers.

8.1 CREATIVE TOURISTS' PREFERENCES

In Chapter 4, we gained an overview of the dimensions that are important to modern creative tourists: safety, convenience, aesthetic appeal, liveliness, and the ability to live like locals. Ongoing studies helped us learn a bit more about these dimensions and include bits that these dimensions may leave out.

First of all, diversity was mentioned a lot, whether it was C2 describing how all the photos only reflected one "place" of the neighborhood or D1, D19, and D20 explicitly looking for a diverse set of photos. We saw this in our initial study in Chapter 4 too, considering "diversity" one of our potential dimensions, until the survey showed people didn't view it as a separate construct. Perhaps the disconnect happened because survey respondents saw the word "diversity" only to mean racial diversity. Interviewees talked about diversity of people, but also diversity of actions; they wanted to see the breadth of possible things they could do there. See Figure 8.1; a place with all young rich white people, say, but lots of things to do is not ideal, as it is seen as homogenous. But a diverse suburban neighborhood would also not be ideal because there is nothing for travelers to do there; this would be the "Residential" quadrant. The ideal location would have both a variety of people and a variety of things to do there.

Another dimension that was not fully included in our model, but that kept being mentioned, was uniqueness. People travel to find "something different, not something the same" (C1). This may have been the reason that neighborhood comparison was not a well-liked way to navigate. We had focused on finding things that were unique within a city, like using Twitter to find words that were popular in one neighborhood but not popular in other neighborhoods. However, a better model might be to find words that are unique between cities, then find where those words are found within cities. Pierogies are more popular in Pittsburgh than in most cities, so finding a

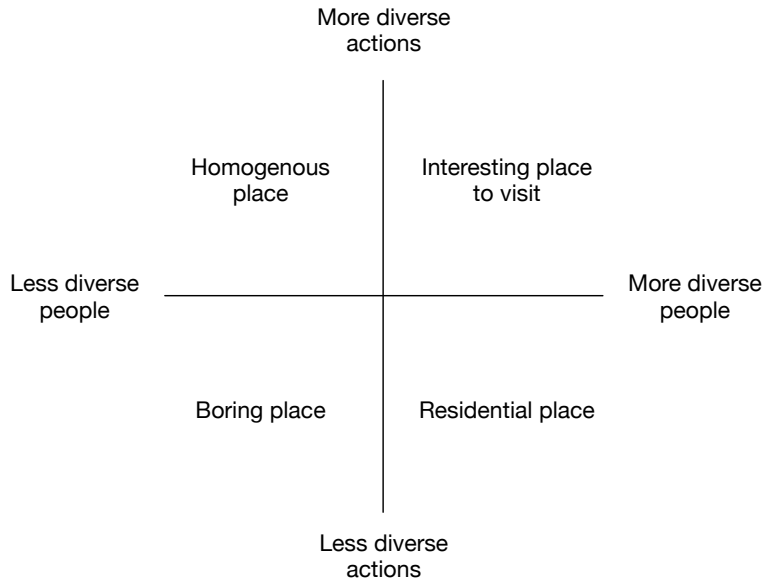


Figure 8.1: Dimensions along which places can be diverse

great Pittsburgh pierogi place, or learning about the pierogi race at Pittsburgh Pirates games, would be a unique and enjoyable experience.

Third, "authenticity" came up in reports from participants like D13 and D20. Participant D20 gave the most succinct definition of "authenticity": something is authentic if it is run for locals, not just to take tourists' money. This may be a valid, limited way to consider authenticity without solving the entire, complex question of authenticity as put forth by authors like MacCannell [56] and Wang [95]. We could create a new term, "Weak" Authenticity, which is satisfied as long as one simply avoids cynical tourist traps.

MacCannell argued that tourists always want to get further backstage, but our research suggests that that is slightly oversimplified. Certainly, participants want to avoid some of the most egregious front stage places, like D13 talking about "staged" photos at tourist attractions or D21 avoiding "gimmicky, touristy" sights. However, sometimes participants like C6 and B7 wanted to be more central (and therefore in a more front stage place) due to language difficulties. They were interested in the daily life of people around where they were traveling, but because they were not fluent in Czech and Spanish, respectively, they had to stick to more front stage locations. (Undoubtedly this would be a wider concern if we had focused more on international travelers.) We can therefore see authenticity as a tension between competing factors: participants want to see more of the backstage of a place, but they are often limited by practical concerns. MacCannell acknowledged this as well, but in our research, we found the practical concerns usually outweighing this search for the "most authentic" places. As a result, our new term of "Weak Authenticity", which is limited to merely avoiding the worst, most inauthentic places, becomes a better criterion in our model than stronger, more complicated versions of authenticity.

8.1.1 Enriching the model with "Traps"

Tourist traps are one class of trap that a traveler might run into. A wider definition of "traps" may be a useful way to refine the model, as it would take into account the way people want to know about our five dimensions. The dimensions are not all equally weighted, nor are they understood linearly. Safety and convenience, in particular, should be seen not as linear attributes, but rather as thresholds or gates. For most participants, if a place was within their comfort zone, they didn't give the safety of it a second thought. If it was "dangerous," though, the area was immediately ruled out. Similarly, a place doesn't have to be as connected as Times Square; it just has to be accessible enough based on where they want to go and their modes of transportation. Instead of five dimensions participants are trying to maximize, there may be a few thresholds and a few positive aspects. See Figure 8.2 for a potential third iteration of this model, showing the three thresholds and four positive aspects of neighborhood search.

Of course, more research is needed to further develop the model of creative tourist needs and confirm our new model. The way to integrate the dimensions of diversity and uniqueness into our model is an open question. Tourism researchers, too, should further define what constitutes a "creative" tourist, as opposed to a "cultural" or "sun and sand" tourist. Furthermore, these definitions should include some concept of tourists being multiple types; few people are singularly one type or another. Even within the same trip, someone may spend one day seeing the sights and the next day trying to blend in as a local.

8.2 How social media and other data can help these tourists

With these updated preferences in mind, we return to the application question: how can we build something to help tourists learn what they need to know? Based on our studies, we would provide the following recommendations for future tourist guide developers.

8.2.1 Prior work and improvements

Others have researched ways to help tourists learn about cities they visit. As described in Chapter 2, many of these have focused on building recommendation systems (e.g. [49, 84]), which does not adequately address the human needs involved in tourism. In addition, tools like Yelp and TripAdvisor help people learn about what businesses exist in a given place, but they are too mechanical to give tourists a thorough picture. These sites focus on the space, while the work in this document aims to give users a sense of the place [30].

Projects like Curated City [17] and Where is the Soho of Rome [51] try to give people a sense of the place through different means. Curated City lets people contribute what they love about different neighborhoods; however, I avoid the substantial bootstrapping problem of such a system by using data that people are already posting. Neighborhood comparison as in [51] gives a qualitative sense of neighborhoods by comparing them to others, but the output is not very rich. Neighborhood Guides can be thought of as a browsing tool, a way for readers to get a rich image of a neighborhood and interact with data in different forms.

Having built such a tool, I will spend the rest of this chapter providing recommendations for others building similar tools.

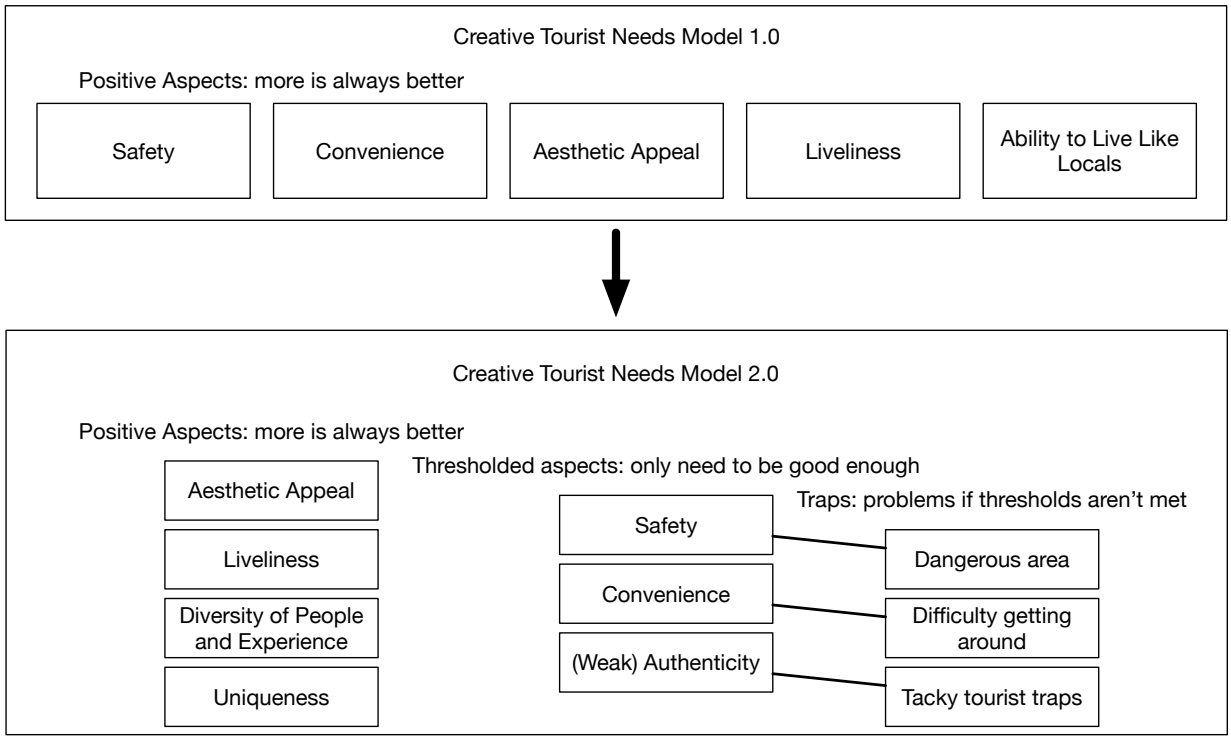


Figure 8.2: A potential next version of the creative tourist information search model. Instead of five undifferentiated aspects, we realize that some aspects are to be maximized while others only need to be "good enough."

8.2.2 SIMPLIFY MAPS, ADD SURPRISE, AND INTEGRATE MORE DATA

A map remains an important navigation point. It is the quickest, easiest way to understand the basic layout of a city. Given that tourists have to get around an unfamiliar place, it is critical that they have at least a rough idea of where these places are. In addition, when we overlaid data on top of the map, participants understood it much more quickly and easily. As we discussed in Chapter 7, most statistics are simply proxies for residential density, so we could replace most of our map with a density map, adding a crime map if needed.

Correll and Heer [13] have noticed a similar problem on large maps, like maps of US counties or Canadian provinces: most maps of these areas are simply population density maps. Even maps that are corrected for population density (by using per-capita statistics instead of absolute statistics) fail because they instead highlight sparsely populated areas with large relative (and mostly random) fluctuations. They describe a Bayesian algorithm used to generate "Surprise Maps" that show how surprising an area is. This would fit perfectly with our needs: instead of showing crime rates, for example, the map could show only areas that have surprisingly high crime, helping visitors avoid that area.

More obviously, it would be helpful to integrate chloropleth maps with other data sources. Many times during our studies, participants would say something like, "this area looks interesting. What hotels are there?" We would have no way to answer them. Integrating sites like Booking.com, Airbnb, and Yelp would help people find not only what areas they like, but also what specific things in those areas. Some of these maps could be quite detailed; as D15 said, she didn't want an area that just has "shopping" as much as one that has a specific kind of shopping. Even more useful functionality would enable users to save some spots while looking at a map, so they can pick out the sights they want to see or other personally-relevant venues, then see how accessible those places are from a number of potential neighborhoods.

8.2.3 SELECT THE BEST PHOTOS

Photos remain our most useful single data source, as each one can convey a lot more information, and trigger more imagination, than simple words can. However, we learned many ways to improve the information that people can glean from a set of photos. This involves selecting the right photos and presenting them in the right way.

There are many ways to select the "right" photos to describe an area. We reimplemented one, the algorithm in [37], but did not find that it was significantly more useful to our participants than random photos. We chose this method over other approaches, because they were all somewhat unsuitable for our particular task. For example, Crandall et al [15] selected a "representative" image for each city and each landmark, but we wanted more than one photo so users could explore an area themselves. Other approaches looked to first understand the "representative tags" for an area [4, 43, 64, 73], but as our tags came from a closed vocabulary based on a computer vision system, the same approach may not apply. Some, like [43], went on to suggest representative images based on these tags, but these were often based on images of the same landmark; "representativeness" for them means it's a valid picture of that landmark, while for us it means it shows life in the neighborhood well.

Our work, however, suggests that automatizing the process too much can have its own pitfalls. First, it can focus too much on one side of the city, as when it only showed runners or only showed white and Asian people. Second, an issue of trust appears: can users trust that the photos are a fair representation of a place if a complex algorithm is selecting them? As a result, we believe that the best way forward is to apply a few simple algorithmic filters to weed out "bad" photos, then to randomly select from what remains.

Some of these filters could be based on the autotags Flickr provides. First, remove photos that only have faces, or those that only have buildings. Other heuristics could be developed as well, like only selecting photos with at least some people in them, or only taking outdoor photos. Non-autotag-based heuristics should apply as well, like taking photos from multiple photographers and multiple days, so that the entire photo set is not one viewpoint or one event. Finally, some notion of image quality should be included, too; perhaps if the photo is too bright, dark, or blurry, leave it out.

In summary, we suggest the following heuristics for picking photos that will be helpful for travelers:

1. Select photos of people doing things.
2. Do not select too many photos from the same person.
3. Do not select too many photos from the same day.
4. Select photos with good overall quality (lighting, sharpness).
5. Sort photos somehow, so they are not overwhelming.

8.2.4 USE MORE FORMS OF STRUCTURED TEXT

While text did not play a large role in our studies, we still think that it would be useful, due to participants' propensity to look for short descriptions of neighborhoods. We would adjust from our approach, though, in a few ways.

First, while tweets are rich sources of data, their free-text nature means that it is difficult to identify key words. A lot of the terms we identified in 5 were venue names; this is not inherently problematic, but can be frustrating if one wanted to learn about other important things besides venues. One way to remedy this might be to select terms from a closed vocabulary. As a test, we ran a simple experiment to find each city's favorite foods on Twitter. We used a closed vocabulary of about 500 names of foods[1], and assigned each one a score for each city based on simple version of the TF-iDF algorithm. In this case, term frequency was defined as the number of times the term appeared in this city, and document frequency was the number of times it appeared in all cities. Results, shown in Table 8.1, seem at a glance to match up with common knowledge about each place. The question "what foods should I try when I'm in a certain city?" is often on travelers' minds, and by constraining our tweets to answer this question instead of the broader question of "what's popular here?", we are able to give people some more useful information.

[1] Available at https://github.com/dantasse/swot_perderder/blob/master/foods.txt

	Austin		New York		Pittsburgh		San Francisco	
1	pecan	0.56	breadfruit	1.0	pierogi	0.71	cassava	0.88
2	barbecue	0.49	papaya	0.89	pawpaw	0.47	sourdough	0.74
3	tamale	0.45	artichoke	0.87	coleslaw	0.43	dosa	0.73
4	brisket	0.32	empanada	0.87	mints	0.42	loquat	0.71
5	doughnut	0.30	lox	0.85	boysenberry	0.42	marionberry	0.67

Table 8.1: Each city's favorite foods, according to their tweets. Scores reported are TF-iDF scores; they can also be interpreted as "56% of the mentions of pecans were in Austin." Notice the occurrence of iconic and regional foods, like pierogis in Pittsburgh, sourdough in San Francisco, and lox in New York.

Second, we would aim to include more different types of textual data to answer different types of questions. One example we could use would be things that can be done in an area. Dearman et al [20, 21] have developed ways, based on verb-noun pairs in Yelp reviews, to determine what can be done in an area.

Third, we would ideally include some version of the "blurb" that we described in Chapter 7: a sentence-length or paragraph-length description of an area. This would eliminate tourists' needs to search for other websites to provide a conceptual overview of a neighborhood. Unfortunately, no obvious source exists for this type of data. For one way to create it, we look to crowdsourcing; projects like Curated City [17] suggest that people enjoy talking about what they love about their neighborhood, and may be willing to contribute short descriptions of it.

8.2.5 Overall integration

Using a customer journey map (see Figure 8.3), we can track a person's path through using this tool to get useful insights about the neighborhoods in a city they're traveling to. In doing so, we realize that the guide-booking process is a two way street: users do not simply find a neighborhood and stick with it. Instead, they may go back and forth between guides and sites that tell them where they can book. This can be a particular pain point, if they find an area they like, but can't find a place to stay there. As a result, it would be ideal to integrate this tool with sites like booking.com, hotels.com, and Airbnb to help users decide where to stay without any task-switching.

Another potential opportunity is a higher-level city guide, to show the best parts of the whole city, not just each neighborhood. This would then be useful for a trip that's less structured than the one in Figure 8.3, in which the traveler does not already know their destination. They may find that the entire feel of Pittsburgh is appealing to them, or that they might prefer to visit another city. This is particularly relevant in international travel, where culture can differ widely between cities and countries.

Finally, a third opportunity appears once a user arrives at their destination. In addition to showing

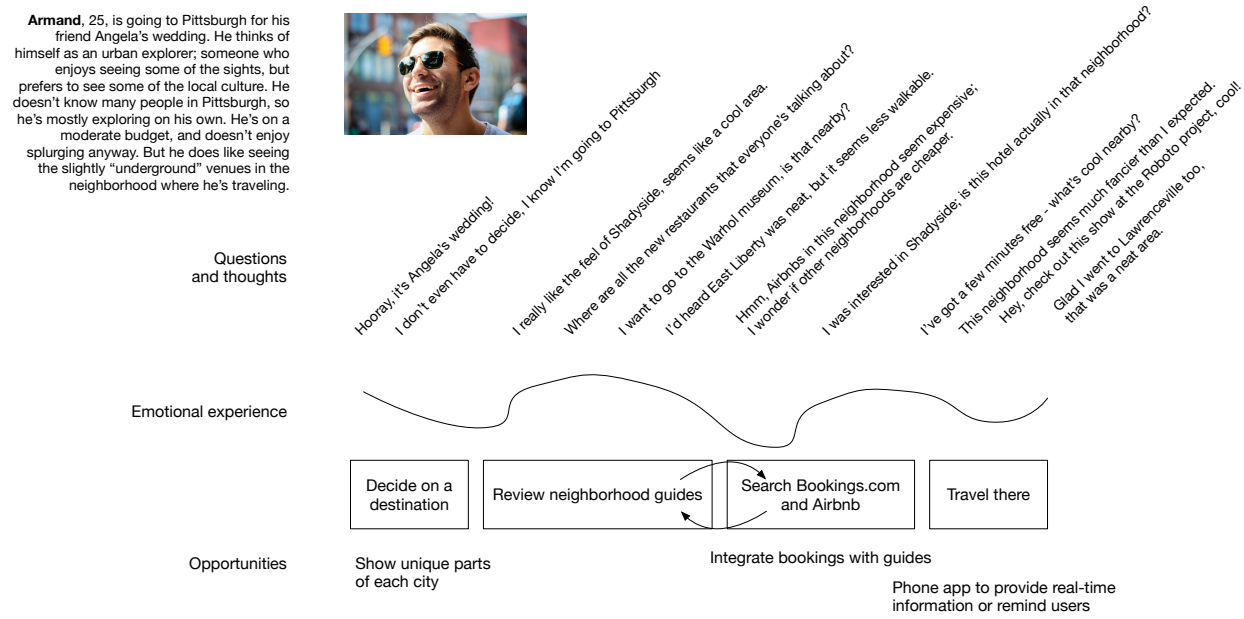


Figure 8.3: A customer journey map showing one customer's hypothetical journey through using this tool.

users useful facts about neighborhoods before they travel, it would be worthwhile to show them local interesting things once they got there. However, this would be a difficult balance; because these travelers mostly want to get outside and experience the city they are in, it's important not to keep them stuck to their phones.

8.3 WHAT SOCIAL MEDIA REVEALS ABOUT OUR CITIES

The question of what public geotagged social media reveals about our cities may be of broader interest to the research community. Many researchers have attempted to use social media to find out factual information. For example, the LiveHoods project used Foursquare check-ins to better understand neighborhoods that reflect social movement within a city [18], and Kafsi et al used Flickr to find tags that are particularly representative of a region [42].

Most of these applications bottom-up or data-driven. The work by Kafsi et al, for example, starts from Flickr data and shows what that data can tell us about the city. Similarly, LiveHoods reveals what Foursquare can tell us about a city.

In this document, I expand on their work, but also connect them to a top-down need: that of the tourist. In doing so, I also help to understand the possible application space. Instead of simply finding out one more thing from the data, I show clues where social media data is useful and where it is not. Particularly, social media data is useful to tell us about the best side of a city, and unhelpful to tell us about quantitative demographics or other more objective information.

8.3.1 WHY NOT TO USE SOCIAL MEDIA QUANTITATIVELY

Geotagged social media is an inherently biased data source because it depends on its user base, which is a small subset of people. Researchers constantly note this, but then often brush it aside or ignore it, because there is no easy way to correct for it.

Sometimes these studies attempt to do something quite precise, like find where users live. Because of the noise in social media data, it is hard to be very precise; results include locating 49.9% of people within 100 miles [31] or 79.9% of people within 200 miles [77]. Other times, they take a broader look at mobility and find patterns like Levy flights in people's movement [7], which is useful in the aggregate but can be difficult to apply to any one individual situation.

Furthermore, the data stream is drying up. As we showed in Chapter 3, the volume of geotagged tweets at the sub-city level is shrinking, and data sources like Instagram are shutting off their public APIs. Anything that attempts to find something quantitative will likely lead to other researchers building on top of it, which is problematic if it's based on a small and shifting data source.

8.3.2 WHAT SOCIAL MEDIA CAN TELL US

However, we have shown that public geotagged social media data does display a useful side of each neighborhood and city. It can be used to show people around a neighborhood, and furthermore to show not just the factual buildings-and-roads view of a neighborhood, but also the subjective "best side" of that neighborhood. Social media posts can show travelers what people do locally, what they like, and what they think is worth recording. Based on our research, we suggest the following principles when trying to interpret geotagged social media data:

- Remember that this data is from a small sample of people.
- Remember that this data is probably consciously geotagged.
- Use the data in such a way that a human can interpret it (i.e. the output should be something for a human to browse, not a number).
- Use it to spark ideation, not confirmation. If you generate a hypothesis from this kind of data, confirm it with more substantial data.
- Do not use it for anything mission-critical. Realize that Twitter, Instagram, and other platform owners can remove it at any time.

Keeping these principles in mind will ensure that the application gains the benefits of the diversity and richness of public geotagged social media data, while not falling prey to many of its pitfalls.

9 CONCLUSION

Travelers are not like they used to be. In the past, tourism sites could put up a list of sites and consider their work finished. Now, though, a new breed of tourists are more demanding: creative tourists want to get out and explore different parts of the city and learn what some of the everyday life around there is like. What does this mean, though; what do these tourists really want? And how can they find it?

At the same time, we are now over a decade past the introduction of Twitter, Flickr, and Facebook, so we have had plenty of time for users to create and share public geotagged data. The question remains, what can we do with that data?

In this thesis, we have addressed these three questions. Through a series of interviews and surveys, we developed an initial answer to the first question, what creative tourists really want. We've found that they want a place that is safe, convenient, lively, and aesthetically pleasing, and that allows them to live like locals wherever they are. We further refined this model to suggest the inclusion of diversity, authenticity, and uniqueness, but further research is needed to refine this revised model.

We have developed a series of answers to the question of how travelers can get the information they need as well, through a series of prototype-based user studies and a quantitative Mechanical Turk study. These led us to the conclusions that there is room for a neighborhood guide application that would help them learn about neighborhoods, and that social media photos are the most useful data source to help them get there.

Finally, we have added to the literature about what social media can tell us about our cities. We provide evidence that social media, instead of being a quantitative data source like the U.S. Census, is best understood as a qualitative data source that people can use to explore new places. We hope to encourage further research in this vein, and further applications that use the richness of social media to aid exploration.

We hope that this encourages more research on social media and more tools for travelers that meet modern travelers' needs. Tourism can be a way for people to relax and have fun, but also a powerful force for helping people understand different people and different cultures, especially in the age of shared lodging sites like Airbnb and Couchsurfing. We hope that the abundance of social media makes this cultural exchange more feasible.

BIBLIOGRAPHY

[1] Airbnb: About Us. https://www.airbnb.com/about/about-us, 2016. 4.1

[2] Couchsurfing: Share your life. http://www.couchsurfing.com/about/about-us/, 2016. 4.1

[3] HomeAway, Inc. is the world's leading online marketplace for the vacation rental industry. https://www.homeaway.com/info/about-us, 2016. [Online; accessed 2016-09-19]. 4.1

[4] Ahern, S., Naaman, M., Nair, R., and Yang, J.H.I. World Explorer: Visualizing Aggregate Data from Unstructured Text in Geo-Referenced Collections. In *Proc. 2007 conference on Digital libraries - JCDL '07*, p. 1. 2007. URL http://portal.acm.org/citation.cfm?doid=1255175.1255177. 2.3.2, 6, 8.2.3

[5] Anderson, M. 6 Facts about Americans and their Smartphones. *Pew Research Center*, (April), 2015. URL http://www.pewresearch.org/fact-tank/2015/04/01/6-facts-about-americans-and-their-smartphones/. 1

[6] Ashworth, G. and Page, S.J. Urban tourism research: Recent progress and current paradoxes. *Tourism Management*, 32(1):1–15, 2011. URL http://dx.doi.org/10.1016/j.tourman.2010.02.002. 2.1, 2.1.1, 4.7

[7] Barchiesi, D., Preis, T., Bishop, S., and Moat, H.S. Modelling human mobility patterns using photographic data shared online. *Royal Society Open Science*, 2015. URL http://dx.doi.org/10.1098/rsos.150046. 2.3.3, 8.3.1

[8] Beyer, H. and Holtzblatt, K. *Contextual design: defining customer-centered systems.* 1997. 3.5, 3.8.3, 7.2

[9] Bird, S., Klein, E., and Loper, E. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 5.2.5

[10] Bradley, R.A. and Terry, M.E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. URL http://www.jstor.org/stable/2334029. 6.2

[11] Buck, M., Ruetz, D., and Freitag, R. ITB World Travel Trends Report. Tech. rep., 2014. 2.1

[12] Clawson, M. and Knetsch, J.L. *Economics of outdoor recreation*. Johns Hopkins Press, 1966. 7.4.1

[13] Correll, M. and Heer, J. Surprise! Bayesian Weighting for De-Biasing Thematic Maps. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):651–660, 2017. 8.2.2

[14] Cramer, H., Rost, M., and Holmquist, L.E. Performing a Check-in: Emerging Practices, Norms and Conflicts' in Location-Sharing Using Foursquare. In *MobileHCI*. 2011. **??**, 3.7, 3.8.2

[15] Crandall, D., Backstrom, L., Huttenlocher, D., and Kleinberg, J. Mapping the World's Photos. *Proceedings of the 18th International Conference on World Wide Web, Madrid*, pp. 761–770, 2009. URL http://www2009.org/proceedings/pdf/p761.pdf. 2.3.2, 8.2.3

[16] Crandall, D.J., et al. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences of the United States of America*, 107(52):22436–41, 2010. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3012474{&}tool=pmcentrez{&}rendertype=abstract. 3.1

[17] Cranshaw, J., Luther, K., Kelley, P.G., and Sadeh, N. Curated City: Capturing Individual City Guides Through Social Curation. In *CHI*. 2014. 8.2.1, 8.2.4

[18] Cranshaw, J., Schwartz, R., Hong, J.I., and Sadeh, N. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. *ICWSM*, 2012. 2.3.1, 5.1.1, 8.3

[19] Csapó, J. The Role and Importance of Cultural Tourism in Modern Tourism Industry. *Strategies for Tourism Industry - Micro and Macro Perspectives*, p. 33, 2012. 1

[20] Dearman, D., Sohn, T., and Truong, K.N. Opportunities Exist: Continuous Discovery of Places to Perform Activities. In *Proceedings of CHI*, pp. 2429–2438. 2011. 8.2.4

[21] Dearman, D. and Truong, K.N. Identifying the Activities Supported by Locations with Community - Authored Content. In *Proceedings of Ubicomp*. 2010. 8.2.4

[22] Edensor, T. Performing tourism, staging tourism. *Tourist Studies*, 1(1):59–81, 2001. URL http://www.nyu.edu/classes/bkg/tourist/a019896.pdf. 2.1.2, 4.5.4

[23] Füller, H. and Michel, B. 'Stop Being a Tourist!' New Dynamics of Urban Tourism in Berlin-Kreuzberg. *International Journal of Urban and Regional Research*, 38(4):1304–1318, 2014. 1, 2.1, 2.1.2

[24] Gao, Y., et al. W2Go: a travel guidance system by automatic landmark ranking. *Proceedings of the international conference on Multimedia - MM '10*, p. 123, 2010. URL http://dl.acm.org/citation.cfm?id=1873951.1873970. 2.2

[25] Goffman, E. The Presentation of Self In Everyday Life, 1959. 2.1.1

[26] Goldberger, J. and Tassa, T. A Hierarchical Clustering Algorithm Based on the Hungarian Method. pp. 1–15. 5

[27] Goodspeed, R. The Limited Usefulness of Social Media and Digital Trace Data for Urban Social Research. Tech. rep., AAAI Technical Report WS-13-04, 2013. 2.3.4

[28] Guha, S. and Birnholtz, J. Can You See Me Now? Location, Visibility and the Management of Impressions on foursquare. In *MobileHCI*, pp. 1–10. 2013. 3.1, **??**, **??**, 3.7, 3.8.1

[29] Hao, Q., et al. Equip Tourists with Knowledge Mined from Travelogues. *Proc. of the 19th International World Wide Web Conference*, pp. 1–10, 2010. URL papers2://publication/uuid/347A699F-B8BB-48D9-A5C3-9F1CCD02E07F. 2.2, 2.3.2

[30] Harrison, Y. and Horne, J.a. "High sleepability without sleepiness". The ability to fall asleep rapidly without other signs of sleepiness. *Neurophysiologie clinique = Clinical neurophysiology*, 26(1):15–20, jan 1996. URL http://www.ncbi.nlm.nih.gov/pubmed/8657094. 8.2.1

[31] Hau-wen Chang, Dongwon Lee, Eltaher, M., and Jeongkyu Lee. @Phillies Tweeting from Philly? Predicting Twitter User Locations with Spatial Word Usage. *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 111–118, 2012. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6425775. 8.3.1

[32] Hecht, B. and Stephens, M. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *ICWSM*. 2014. 2.3.4

[33] Holtzclaw, J., et al. Location Efficiency: Neighborhood and Socio-Economic Characteristics Determine Auto Ownership and Use - Studies in Chicago, Los Angeles and San Francisco. *Transportation Planning and Technology*, 25(1):1–27, 2002. 4.5.2

[34] Horozov, T., Narasimhan, N., and Vasudevan, V. Using location for personalized POI recommendations in mobile environments. *Proceedings - 2006 International Symposium on Applications and the Internet, SAINT 2006*, 2006:124–129, 2006. 2.2

[35] Jacobs, J. *The Death and Life of Great American Cities*. Vintage Books ed. Vintage Books, 1961. URL https://books.google.com/books?id=P{_}bPTgOoBYkC. 4.5.3

[36] Jacobsen, J.K.S. Anti-tourist attitudes: Mediterranean charter tourism. *Annals of Tourism Research*, 27(2):284–300, 2000. 2.1.2

[37] Jaffe, A., Naaman, M., Tassa, T., and Davis, M. Generating summaries and visualization for large collections of geo-referenced photographs. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval - MIR '06*, p. 89, 2006. URL http://portal.acm.org/citation.cfm?doid=1178677.1178692. 2.3.2, 6, 5, 6.2.1, 6.3.2, 7.1.1, 7.4.2, 8.2.3

[38] Jiang, L., et al. Delving Deep into Personal Photo and Video Search. *WSDM*, 2017. 3.4.2

[39] Johnson, I.L., Sengupta, S., Schöning, J., and Hecht, B. The Geography and Importance of Localness in Geotagged Social Media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2016. 3.8.2

[40] Joseph, K., Tan, C.H., and Carley, K.M. Beyond Local, Categories and Friends: Clustering foursquare Users with Latent Topics. In *Ubicomp*. 2012. 2.3.3

[41] Jurgens, D., Mccorriston, J., and Ruths, D. Geolocation Prediction in Twitter Using Social Networks: A Critical Analysis and Review of Current Practice. In *ICWSM*. 2015. 3.1

[42] Kafsi, M., Cramer, H., Thomee, B., and Shamma, D.a. Describing and Understanding Neighborhood Characteristics through Online Social Media. In *WWW*. 2015. 2.3.2, 6, 8.3

[43] Kennedy, L., et al. How Flickr Helps us Make Sense of the World: Context and Content in Community-Contributed Media Collections Categories and Subject Descriptors. In *ACM Multimedia*. 2007. 2.3.2, 6, 8.2.3

[44] Kinsella, S., Murdock, V., and O'Hare, N. "I'm Eating a Sandwich in Hong Kong": Modeling Locations with Tweets. *Proceedings of the 3rd international workshop on Search and mining user-generated contents - SMUC '11*, p. 61, 2011. URL http://dl.acm.org/citation.cfm?doid=2065023.2065039{%}5Cnhttp://dx.doi.org/10.1145/2065023.2065039. 2.3.2

[45] Komninos, A., Stefanis, V., Plessas, A., and Besharat, J. Capturing Urban Dynamics with Scarce Check-In Data. *Pervasive Computing*, pp. 20–28, 2013. 2.3.3

[46] Krikorian, R. New Tweets per second record, and how!, 2013. URL https://blog.twitter.com/2013/new-tweets-per-second-record-and-how. 1

[47] Krumm, J. Inference Attacks on Location Tracks. *Pervasive Computing*, 10(Pervasive):127–143, 2007. URL http://research.microsoft.com/en-us/um/people/jckrumm/publications2007/inferenceattackrefined02distribute.pdf. 3.9.3

[48] Krumm, J. and Horvitz, E. Eyewitness : Identifying Local Events via Space-Time Signals in Twitter Feeds. *Proceedings of the 23nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2015. 2.3.2

[49] Kurashima, T., Iwata, T., Irie, G., and Fujimura, K. Travel route recommendation using geotags in photo sharing sites. *Proc. 19th ACM international conference on Information and knowledge management*, pp. 579–588, 2010. URL http://doi.acm.org/10.1145/1871437.1871513. 2.2, 8.2.1

[50] Lathia, N., Quercia, D., and Crowcroft, J. The Hidden Image of the City : Sensing Community Well-Being from Urban Mobility. pp. 1–8. 2.3.3

[51] Le Falher, G., Gionis, A., and Mathioudakis, M. Where is the Soho of Rome? Measures and algorithms for finding similar neighborhoods in cities. In *ICWSM*. 2015. 2.3.2, 4.6.1, 5.3, 8.2.1

[52] Leetaru, K.H., et al. Mapping the Global Twitter Heartbeat: The Geography of Twitter. *First Monday*, 18(5):1–12, 2013. URL http://firstmonday.org/ojs/index.php/fm/article/view/4366/3654. 3.1

[53] Lindqvist, J., et al. I'm the Mayor of My House: Examining Why People Use foursquare - a Social-Driven Location Sharing Application. In *CHI*. 2011. URL http://dl.acm.org/citation.cfm?id=1979295. 3.1, **??**, **??**, **??**, **??**, 3.7, 3.8.2, 3.8.3, 5.2.4

[54] Lingel, J., Naaman, M., and Boyd, D. City, self, network: transnational migrants and online identity work. *CSCW*, pp. 1502–1510, 2014. URL http://dl.acm.org/citation.cfm?id=2531693. **??**, 3.7

[55] Liu, W., Zamal, F.A., and Ruths, D. Using social media to infer gender composition of commuter populations. In *AAAI*. 2011. 2.3.3

[56] MacCannell, D. Staged Authenticity: arrangements of Social Space in Tourist Settings. *American Journal of Sociology*, 682(3):678–682, 1977. 2.1.1, 4.5.4, 8.1

[57] Maitland, R. Everyday life as a creative experience in cities. *International Journal of Culture Tourism and Hospitality Research*, 4(3):176–185, 2010. URL http://westminsterresearch.wmin.ac.uk/7203/. 2.1, 4.5.4

[58] Manaugh, K. and El-Geneidy, A. Validating walkability indices: How do different households respond to the walkability of their neighborhood? *Transportation Research Part D: Transport and Environment*, 16(4):309–315, 2011. 5.2.2

[59] McNaught, C. and Lam, P. Using wordle as a supplementary research tool. *Qualitative Report*, 15(3):630–643, 2010. URL http://www.scopus.com/inward/record.url?eid=2-s2.0-77953058705{&}partnerID=tZOtx3y1. 2.3.2

[60] Mohammady, E. and Culotta, A. Using County Demographics to Infer Attributes of Twitter Users. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pp. 7–16. 2014. URL http://www.aclweb.org/anthology/W/W14/W14-2702. 3.1

[61] Naaman, M., Zhang, A.X., Brody, S., and Lotan, G. On the Study of Diurnal Urban Routines on Twitter. *Sixth International AAAI Conference on Weblogs and Social Media*, pp. 258–265, 2012. URL http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4610{%}5Cnhttp://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4610/4991. 2.3.3

[62] O'Connor, B., Krieger, M., and Ahn, D. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*, pp. 384–385. 2010. 5.2.5

[63] of Justice Federal Bureau of Investigation, U.D. Uniform Crime Report, Crime in the United State 2011, Offense Definitions. Tech. rep., 2012. URL https://ucr.fbi.gov/crime-in-the-u.s/2011/crime-in-the-u.s.-2011/11offensedefinitions_final.pdf. 5.2.1

[64] O'Hare, N. and Murdock, V. Modeling locations with social media. *Information Retrieval*, 16(1):30–62, 2013. 8.2.3

[65] Okuyama, K. and Yanai, K. A travel planning system based on travel trajectories extracted from a large number of geotagged photos on the Web. *The Era of Interactive Media*, pp. 657–670, 2013. 2.2

[66] Oldenburg, R. *The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Hangouts at the Heart of a Community*. 1989. 5, 4.5.4

[67] Paldino, S., et al. Urban magnetism through the lens of geo-tagged photography. *EPJ Data Science*, 4(1), 2015. URL http://www.epjdatascience.com/content/4/1/5. 1

[68] Pearce, P.L. and Moscardo, G.M. The Concept of Authenticity in Tourist Experiences. *Journal of Sociology*, 22(1):121–132, 1986. 4.5.4

[69] Prebensen, N.K., Larsen, S., and Abelsen, B. I'm not a typical tourist: German tourists' self perception, activities and motivations. *Journal of Travel Research*, 41(May 2003):416–420, 2003. URL http://jtr.sagepub.com/cgi/content/abstract/41/4/416. 2.1.2

[70] Quercia, D., Ellis, J., Capra, L., and Crowcroft, J. Tracking Gross Community Happiness from Tweets. In *CSCW*. 2011. 2.3.3

[71] Quercia, D. and Saez, D. Mining urban deprivation from foursquare: Implicit crowdsourcing of city land use. *IEEE Pervasive Computing*, 13(May 2012):30–36, 2014. 1

[72] Rattenbury, T., Good, N., and Naaman, M. Towards automatic extraction of event and place semantics from flickr tags. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '07*, p. 103, 2007. URL http://portal.acm.org/citation.cfm?doid=1277741.1277762. 2.3.2

[73] Rattenbury, T. and Naaman, M. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web*, 3(1):1–30, 2009. 8.2.3

[74] Read, M. This Is the Williamsburg of Your City: A Map of Hip America, 2014. URL gawker.com/this-is-the-williamsburg-of-your-city-a-map-of-hip-ame-1460243062. 4.6.1, 5.3

[75] Řehůřek, R. and Sojka, P. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. ELRA, Valletta, Malta, May 2010. http://is.muni.cz/publication/884893/en. 5.3

[76] Richards, G. Tourism Development Trajectories - From Culture to Creativity? *Encontros Científicos - Tourism & Management Studies*, (6):9–15, 2010. 1, 2.1.1

[77] Rout, D., Bontcheva, K., Preoiuc-Pietro, D., and Cohn, T. Where's @wally? A classification approach to geolocating users based on their social ties. *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, (May):11–20, 2013. 8.3.1

[78] Salton, G. and McGill, M.J. Introduction to modern information retrieval. 1986. 5.2.5

[79] Shamma, D.A. The Social Concerns of Geo-Located Rectangles, 2016. URL https://medium.com/@ayman/the-social-concerns-of-geo-located-rectangles-9b361f34811d. 3.9.2

[80] Simioni, J., Oram, J., and Cope, A.S. The Assault on Copenhagen, 2016. URL https://mapzen.com/blog/assult-on-copenhagen/. 3.9.2

[81] Smith, A. Shared, collaborative and on demand: The new digital economy. Tech. rep., Pew Research Center, 2016. URL http://www.pewinternet.org/files/2016/05/PI{_}2016.05.19{_}Sharing-Economy{_}FINAL.pdf. 4.2

[82] Stors, N. and Kagermeier, A. Motives for using Airbnb in metropolitan tourism why do people sleep in the bed of a stranger? *Regions Magazine*, 299(1):17–19, 2015. 2.1.1

[83] Strauss, A., Corbin, J., et al. *Basics of qualitative research*, vol. 15. Newbury Park, CA: Sage, 1990. 4.2

[84] Takeuchi, Y. and Sugimoto, M. CityVoyager : An Outdoor Recommendation System Based on User Location History. *Ubiquitous Intelligence and Computing*, 4159(Figure 1):625–636, 2006. URL http://www.springerlink.com/content/3l282rm6u8278565/abstract/. 2.2, 8.2.1

[85] Tang, K.P., et al. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In *CHI*, vol. 12, pp. 85–94. 2010. URL http://dl.acm.org/citation.cfm?id=1864363. 3.1

[86] Tasse, D., Liu, Z., Sciuto, A., and Hong, J.I. State of the Geotags: Motivations and Recent Changes. In *International Conference on Web and Social Media (ICWSM)*. 2017. 3

[87] Tasse, D., Sciuto, A., and Hong, J.I. Our House, in the Middle of Our Tweets. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM)*, pp. 691–694. 2016. 3.9.3

[88] Thomee, B. and Rae, A. Uncovering Locally Characterizing Regions within Geotagged Data. In *WWW*. 2013. 2.3.1

[89] Thomee, B., et al. The New Data and New Challenges in Multimedia. *arXiv preprint arXiv:1503.01817*, pp. 1–7, 2015. 1, 4.6.3, 5.2.3, 6

[90] Thomee, B., et al. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59(2):64–73, 2016. URL http://doi.acm.org/10.1145/2812802. 3.1, 3.3

[91] Toyama, K., Logan, R., and Roseway, A. Geographic location tags on digital images. *Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA '03*, (November):156, 2003. URL http://portal.acm.org/citation.cfm?doid=957013.957046. 2.3.2

[92] Tsai, J.Y., Kelley, P.G., Cranor, L.F., and Sadeh, N. Location-Sharing Technologies: Privacy Risks and Controls. *A Journal of Law and Policy for the Information Society*, 6:119–151, 2010. 3.6.3, 3.9.2

[93] Tulsiani, N. and Era, N. What's in a Word? Using location context to build better search tools for travel, 2016. URL http://airbnb.design/whats-in-a-word/. 1

[94] Wakamiya, S., Lee, R., and Sumiya, K. Crowd-sourced Cartography: Measuring Socio-cognitive Distance for Urban Areas based on Crowd's Movement. *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 935–942, 2012. 2.3.1

[95] Wang, N. Rethinking authenticity in tourism experience. *Annals of Tourism Research*, 26(2):349–370, 1999. 2.1.1, 4.5.4, 8.1

[96] Yannopoulou, N., Moufahim, M., and Bian, X. User-Generated Brands and Social Media: Couchsurfing and Airbnb. *Contemporary Management Research*, 9(1):85–90, 2013. URL http://www.cmr-journal.org/article/view/11116. 2.1.1

[97] Zhang, A.X., Noulas, A., Scellato, S., and Mascolo, C. Hoodsquare: Modeling and Recommending Neighborhoods in Location-based Social Networks. In *SocialCom*, pp. 1–15. 2013. 2.3.1, 5.1.1