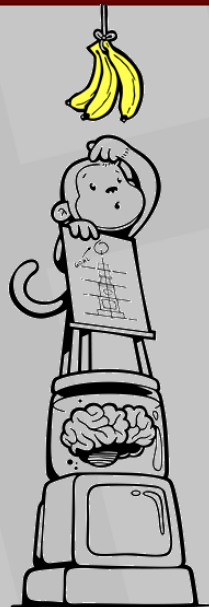


# Social Media and Urban Analytics

American Planning Association  
Oct 20, 2015



• Human-  
Computer  
Interaction  
Institute



**Computer  
Human  
Interaction:  
Mobility  
Privacy  
Security**

**Jason Hong**  
**@jas0nh0ng**

Carnegie Mellon

# Smartphones are Pervasive



- 75% penetration in the US as of late 2014
- About 1.7M Android and iOS apps
- Over 85 billion apps downloaded on each of Android and iOS



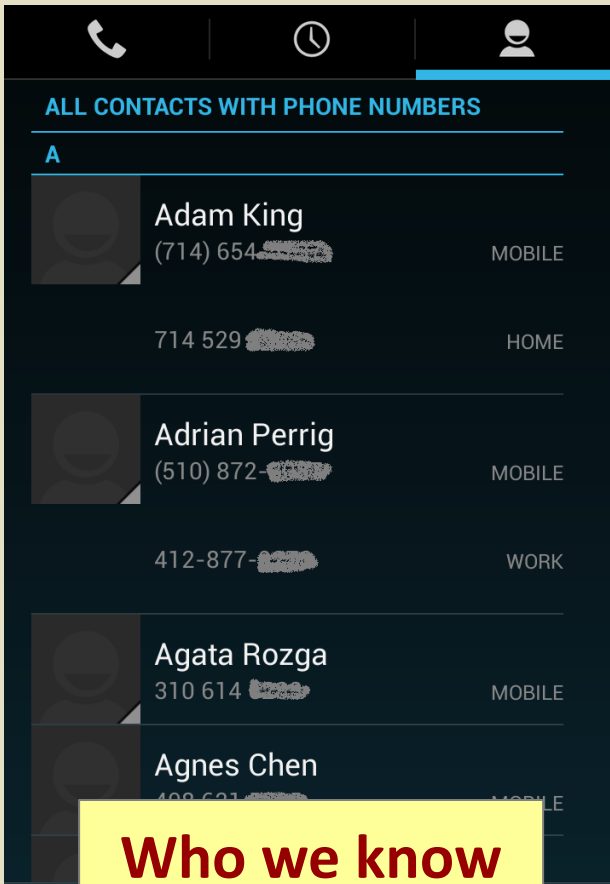
# Smartphones are Intimate



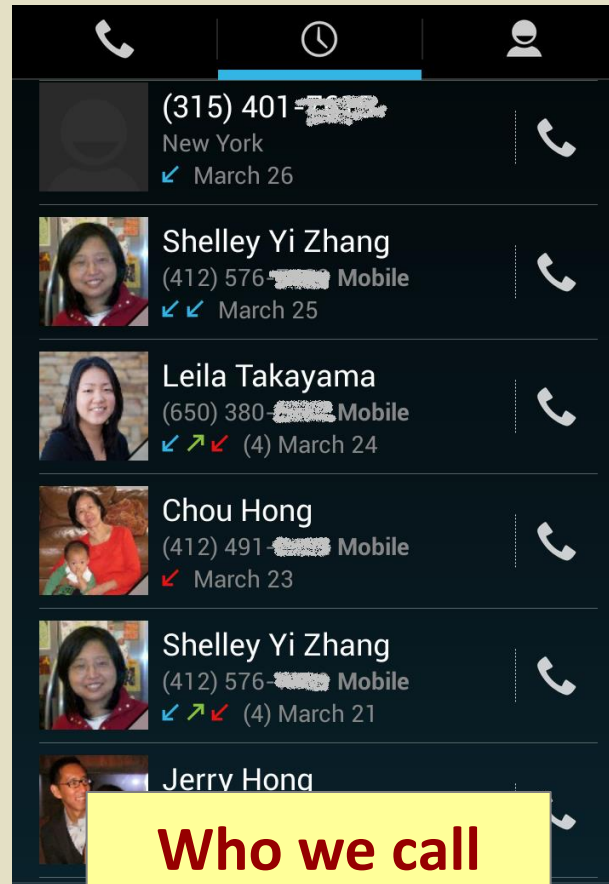
- Mobile phones and millennials (Cisco 2012):
  - 75% use in bed before sleep
  - 83% sleep with their phones
  - 90% check first thing in the morning
  - A third use in bathroom (!!)
  - A fifth check every ten minutes



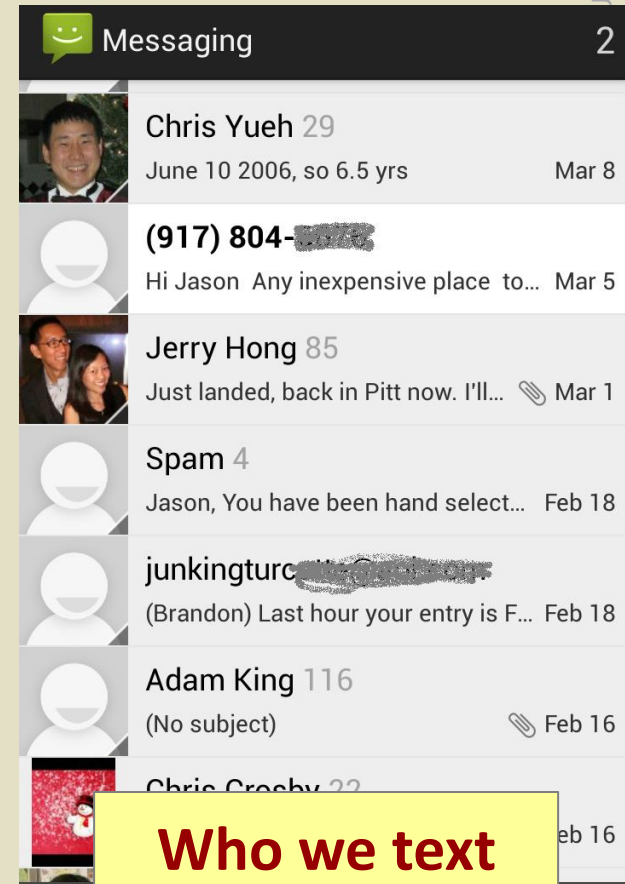
# Smartphone Data is Intimate



**Who we know  
(contact list)**



**Who we call  
(call log)**



**Who we text  
(sms log)**





# Smartphone Data is Intimate



**Photos**  
(some geotagged)



**Where we go**  
(gps, foursquare)



**Sensors**  
(accel, sound, light)



# The Opportunity



- We are creating a worldwide sensor network with these smartphones
- We can now capture and analyze human behavior at unprecedented fidelity and scale



# The Challenge of Getting Data About Our Cities



- Today's methods for getting city data slow, costly, limited
  - Ex. Travel Behavioral Inventory
  - US Census 2010 cost \$13b
  - Quality of life surveys
- Emerging approaches:
  - Installing sensors / cameras
  - Call Data Records



# Understanding Urban Areas

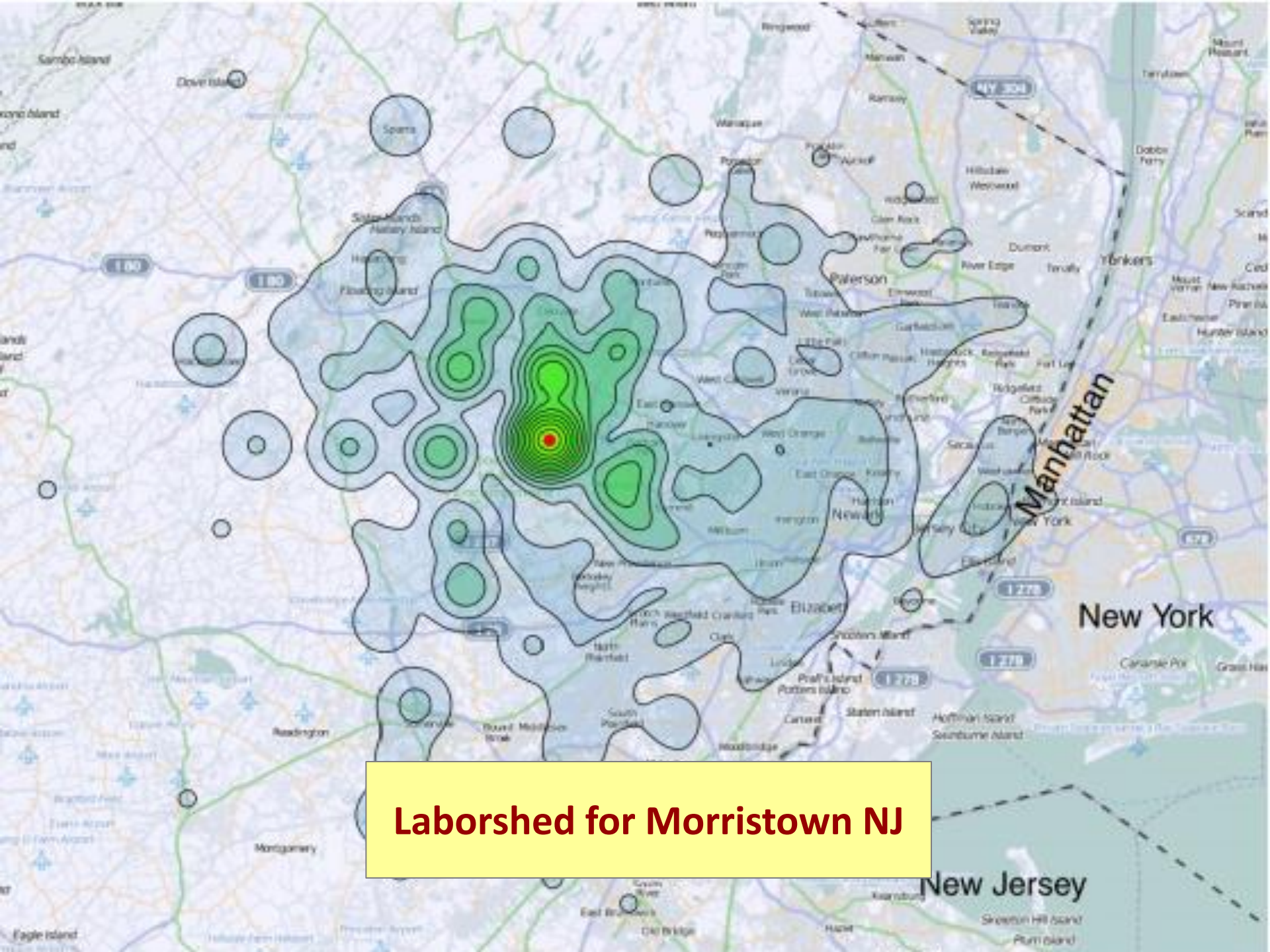
- AT&T Work on Human Mobility



**Median distance  
traveled per day**







**Laborshed for Morristown NJ**

# Other Ways of Gathering Data?

- Call Data Records proprietary
  - No easy access
  - Hard to replicate
- Social Media is an alternative
  - Instagram: 80M photos per day
  - Twitter: 500M tweets per day
  - Foursquare/Swarm: 3-5M check-ins per day
  - Flickr: 1.6M photos per day
  - Small but non-trivial percent is geotagged







**Eric Fischer's Maps of  
Tourists vs Locals**



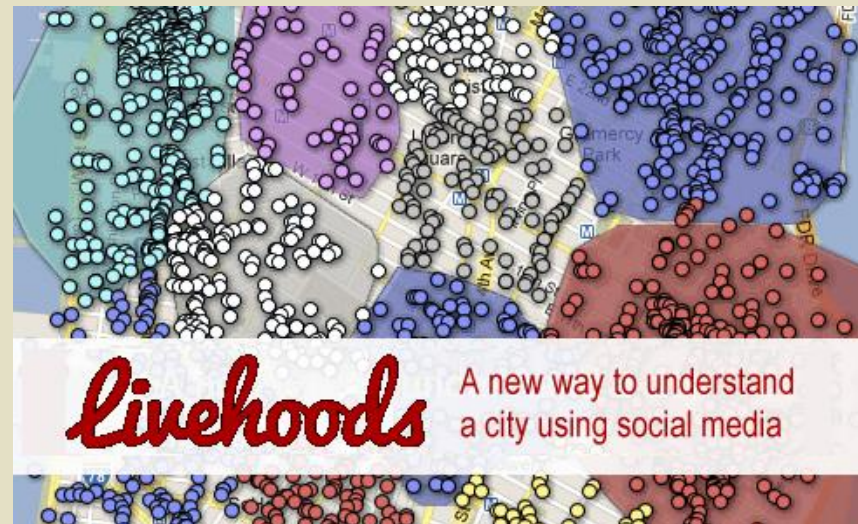
# The Vision: Urban Analytics

- How can we use smartphones + social media + machine learning to offer new and useful insights about cities in a manner that is cheap, fast, and scalable?



# Livehoods, Our First Urban Analytics Tool

- The character of an urban area is defined not just by the types of places found there, but also by the people that make it part of their daily life



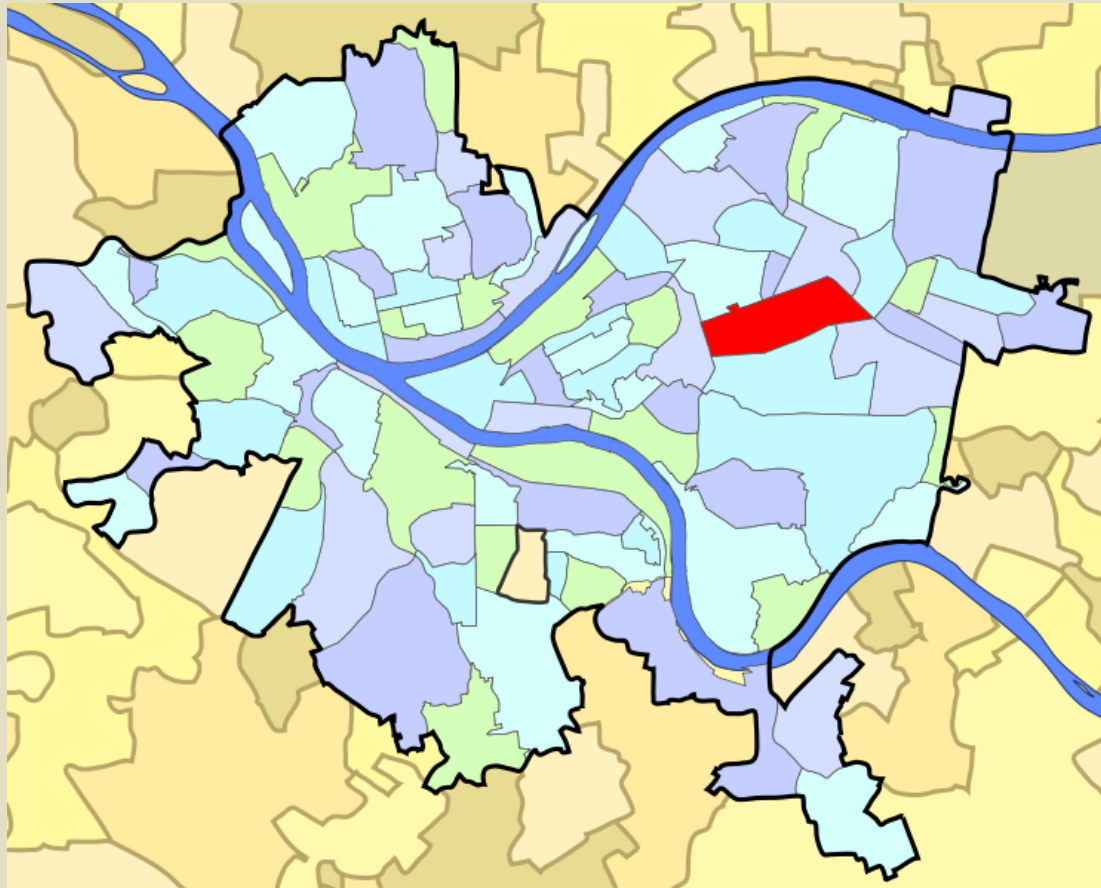
Cranshaw et al, The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City, ICWSM 2012.

**What comes to mind when you picture your neighborhood?**



# The Image of a Neighborhood

You're probably not imagining this.





# The Image of a Neighborhood

What you're imagining probably looks a lot more like this.



*Every citizen has had long associations with some part of his city, and his image is soaked in memories and meanings.*

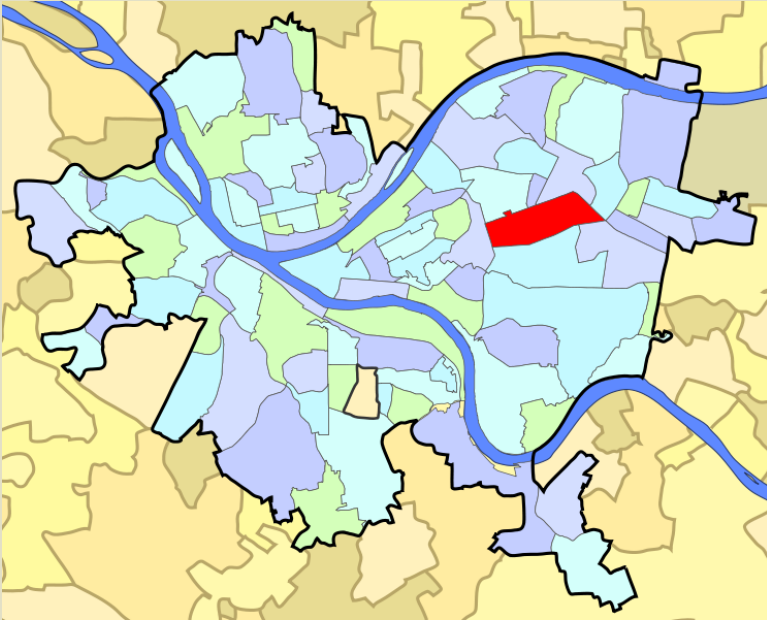
---Kevin Lynch, *The Image of a City*





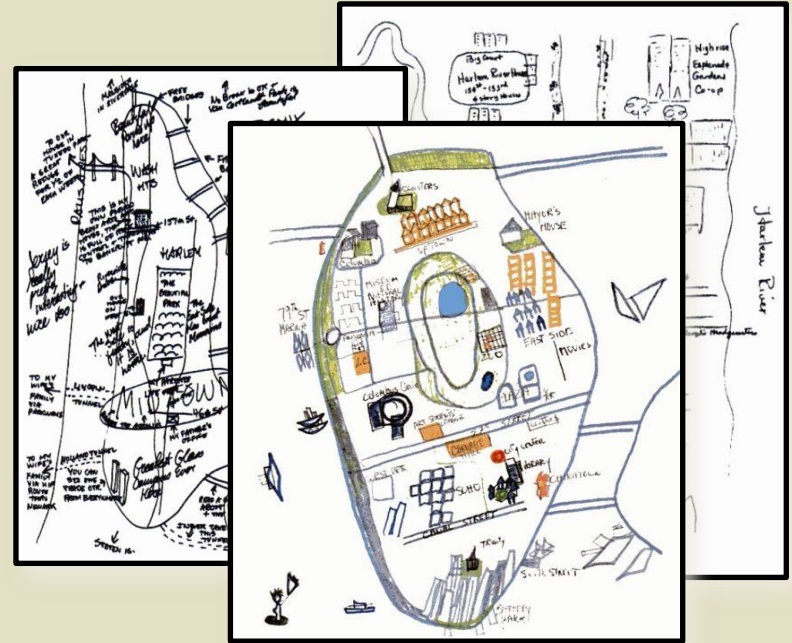
# Two Perspectives

“Politically constructed”



Neighborhoods have fixed borders defined by the city government.

“Socially constructed”



Neighborhoods are organic, cultural artifacts. Borders are blurry, imprecise, and may be different to different people.



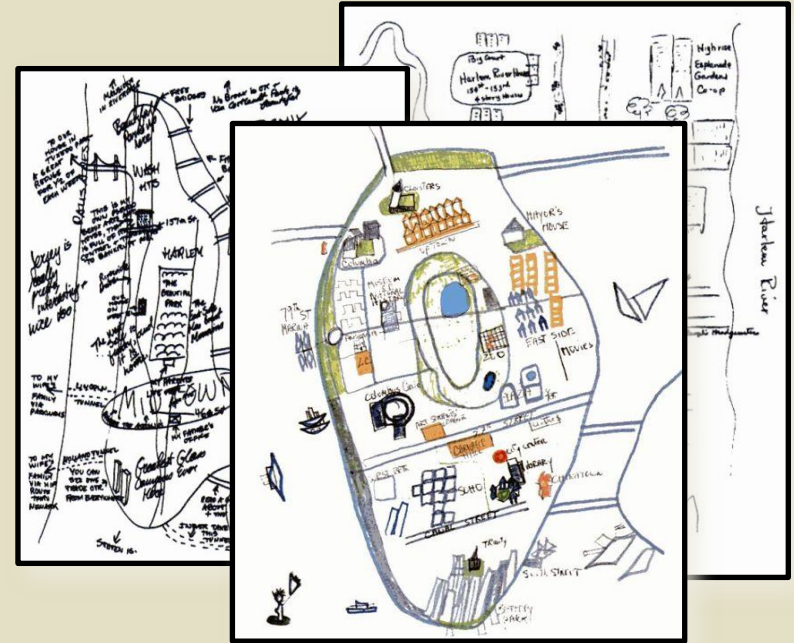
# Two Perspectives of Cities

Can we discover automated ways of identifying the “organic” boundaries of the city?

Can we extract local cultural knowledge from social media?

Can we build a collective cognitive map from data?

“Socially constructed”

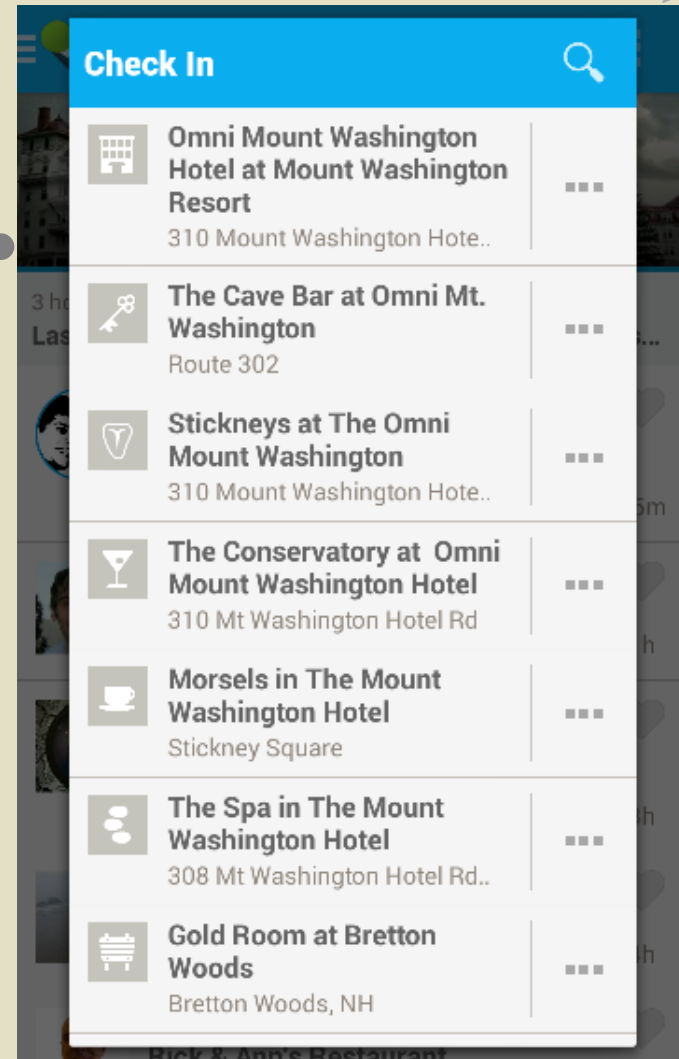


Neighborhoods are organic, cultural artifacts. Borders are blurry, imprecise, and may be different to different people.



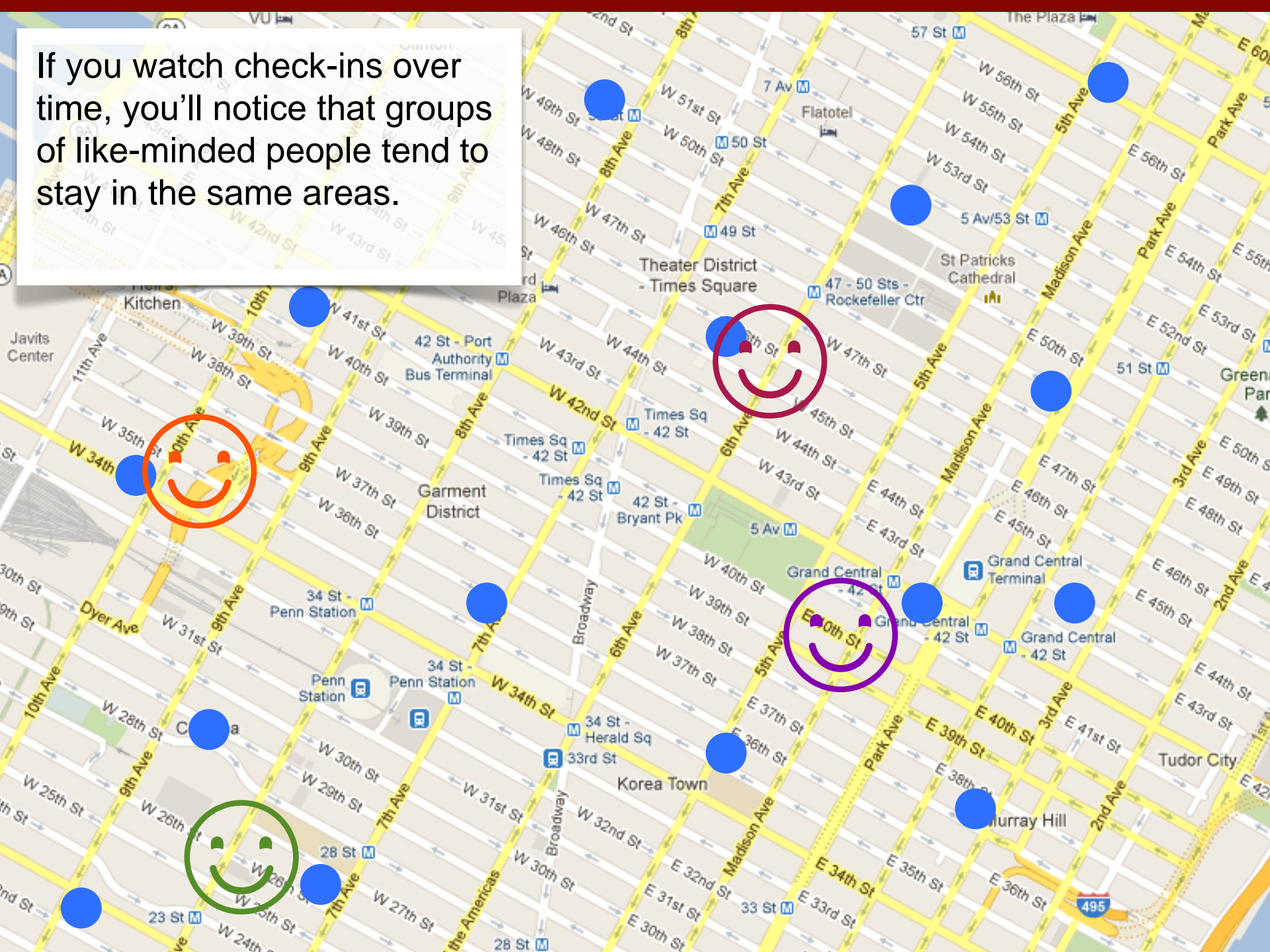
# Livehoods Data Source

- Crawled 18m check-ins from foursquare
  - People who linked their foursquare accts to Twitter
- Spectral clustering based on geographic and social proximity



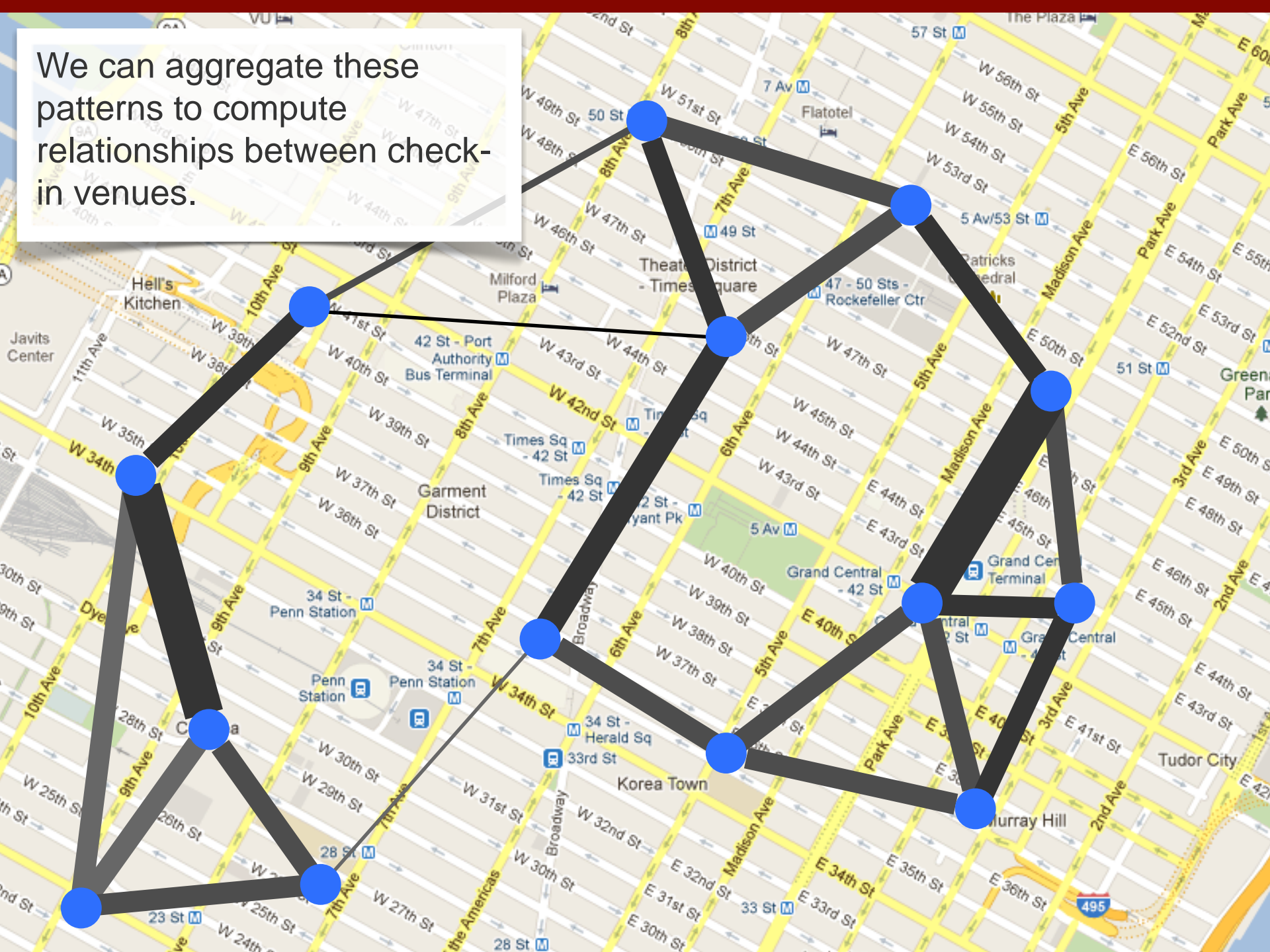


If you watch check-ins over time, you'll notice that groups of like-minded people tend to stay in the same areas.



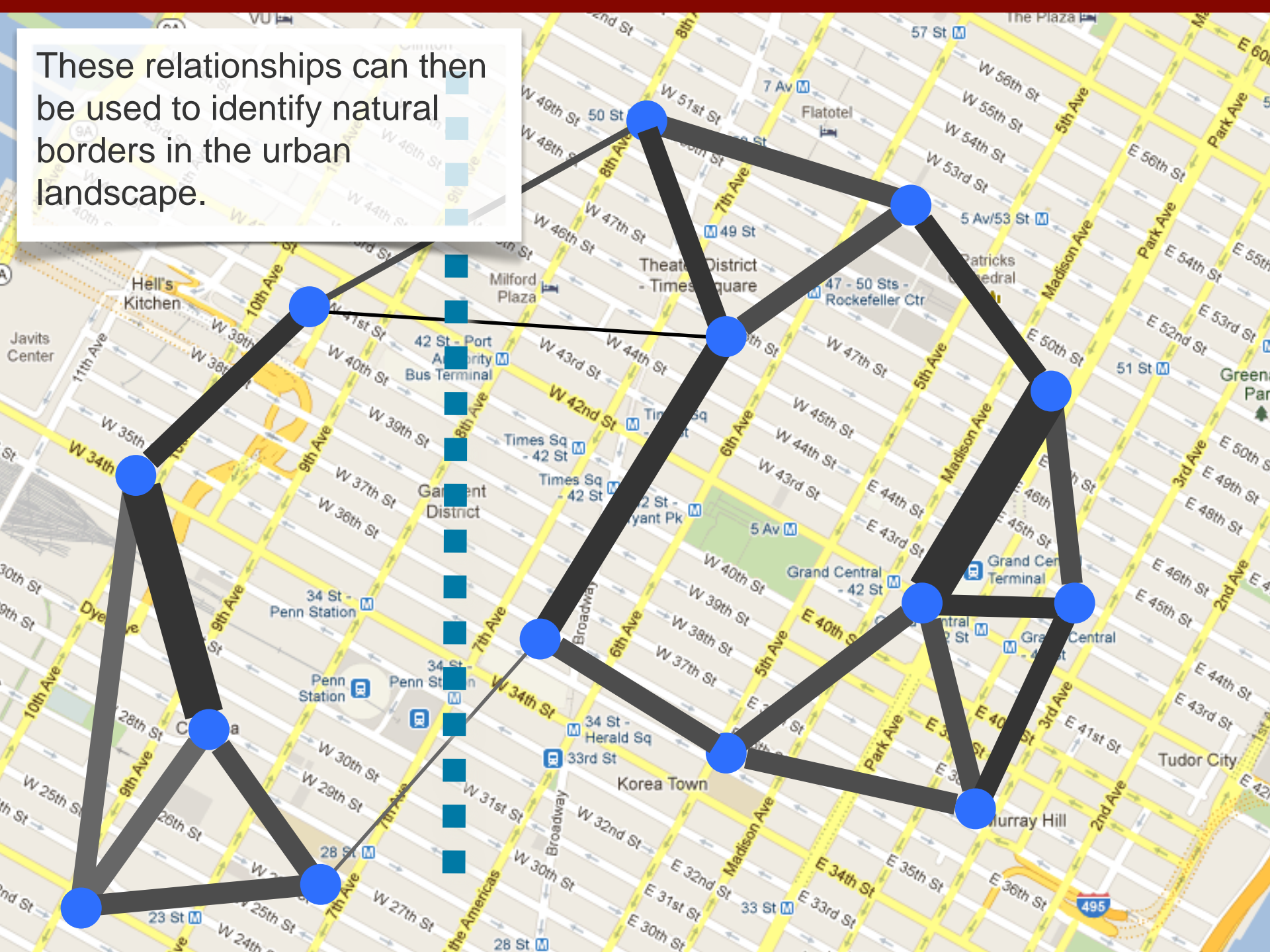


We can aggregate these patterns to compute relationships between check-in venues.



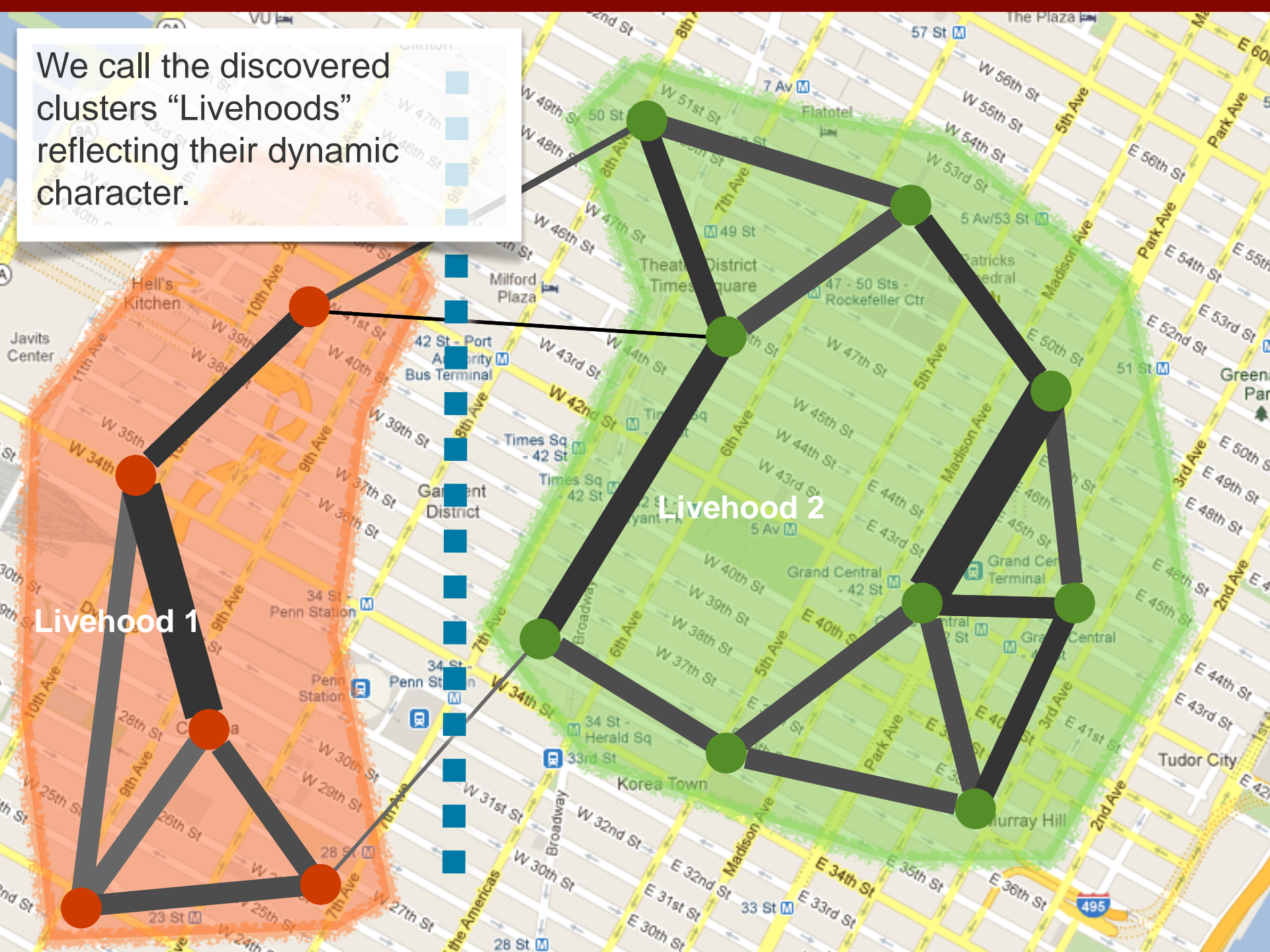


These relationships can then be used to identify natural borders in the urban landscape.





we call the discovered clusters “Livehoods” reflecting their dynamic character.





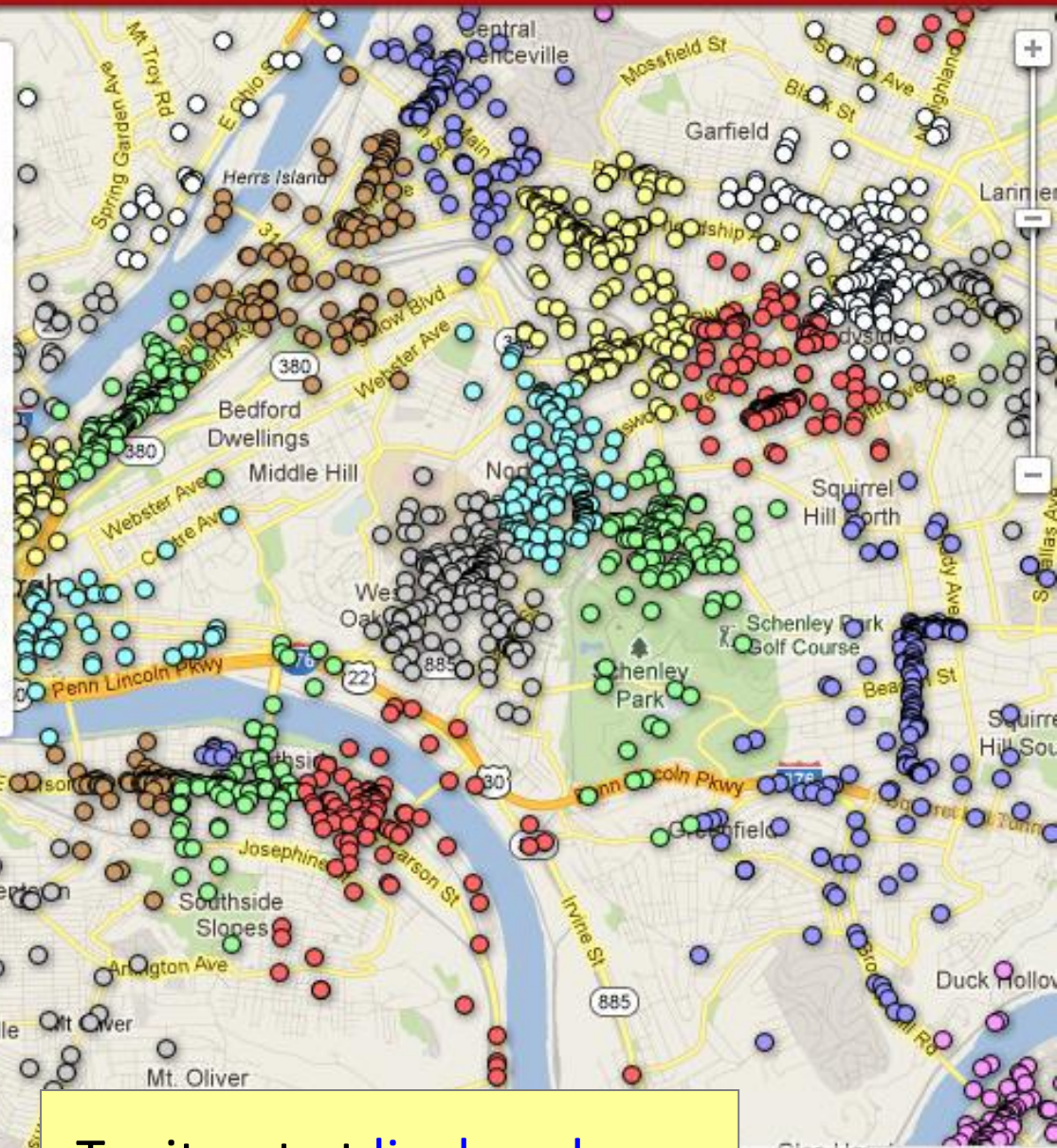
## Welcome to Livehoods!

Each dot on the map (●) represents a check-in location. Groups of nearby dots of the same color form a Livehood.

The shapes of Livehoods are determined by the patterns of people that check-in to them. If many of the same people check-in to two nearby locations, then these locations will likely be part of the same Livehood.

Livehoods reveal how the people and places of a city come together to form the dynamic character of local urban areas.

Click on a location to learn about its Livehood.



Try it out at [livehoods.org](http://livehoods.org)



# Shadyside

## Livehood #41

Character

Related

Stats

Aggregate check-in statistics by day, hour, and type of place reveal usage patterns of the Livehood.

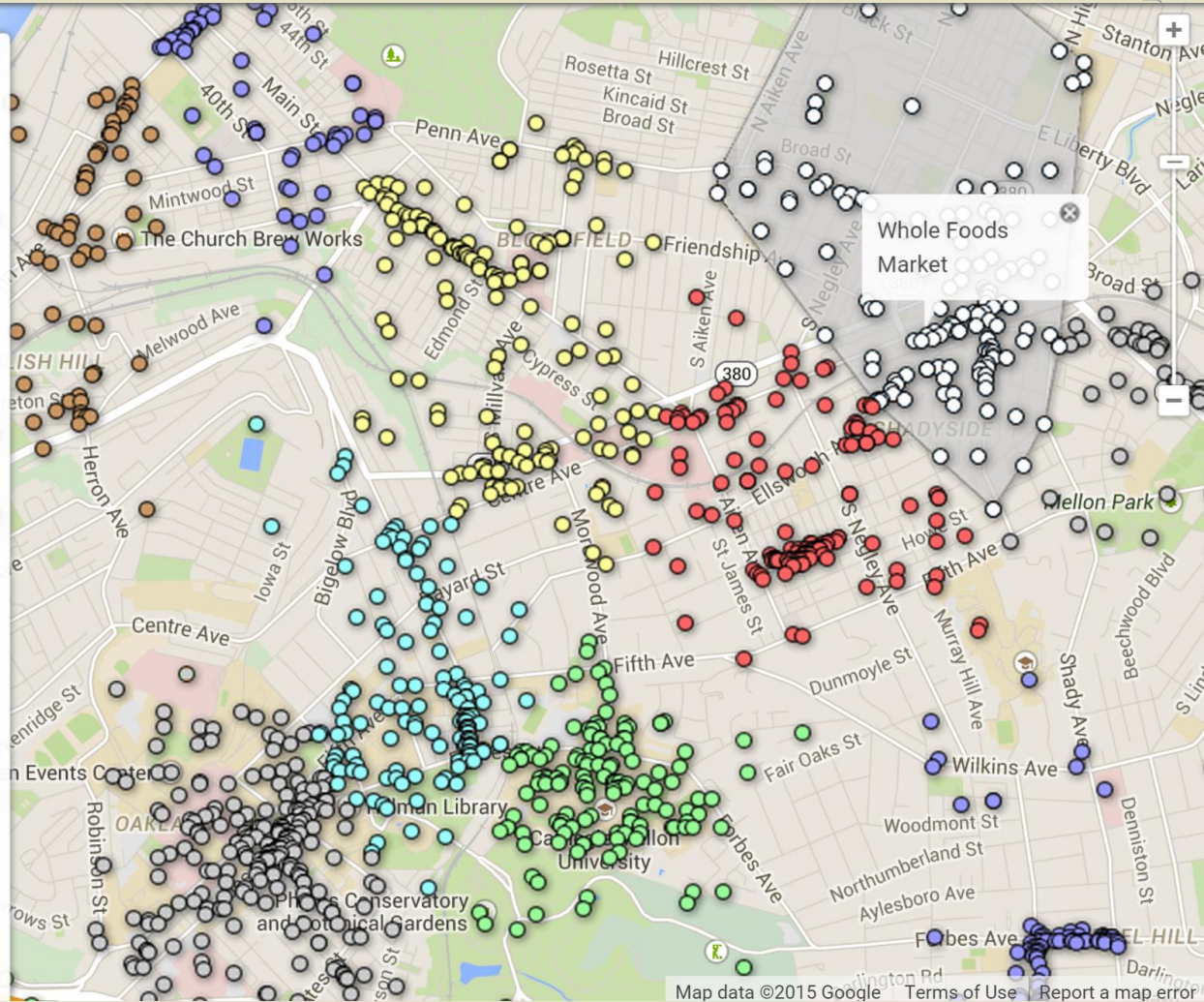
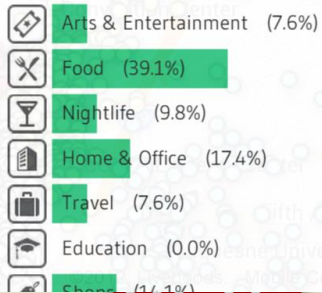
### Daily Pulse



### Hourly Pulse



### Composition



Map data ©2015 Google Terms of Use Report a map error





# PNC Park

## Livehood #42

Character

Related

Stats

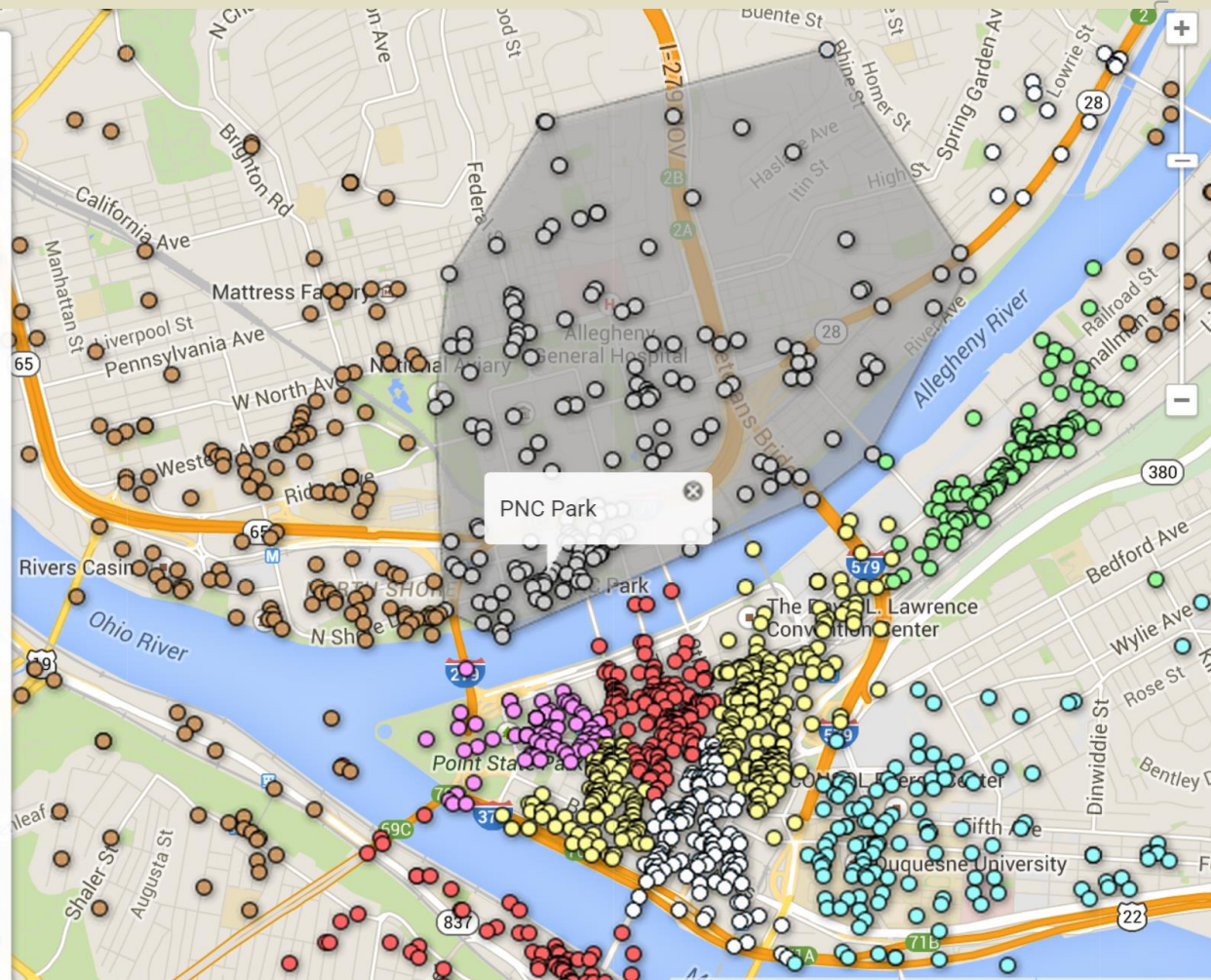
The popular check-in locations and the unique types of places found in a Livehood teach us about its character.

### Top five popular places

- 1 PNC Park
- 2 Allegheny General Hospital
- 3 Andy Warhol Museum
- 4 McFadden's
- 5 Mullens Bar & Grill

### Top five unique things to do here

- 1 Baseball Stadium »
- 2 Building »
- 3 Home »
- 4 Pub »
- 5 Museum »



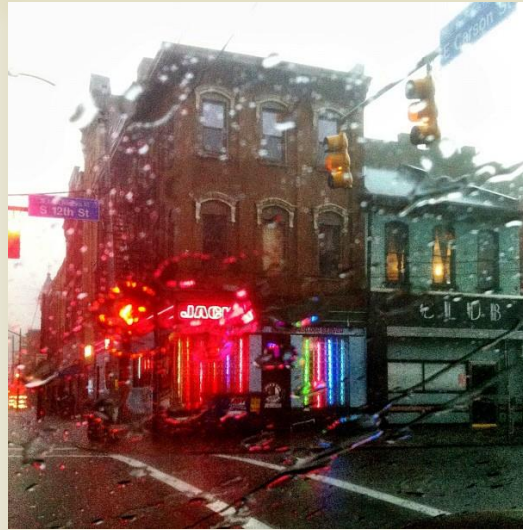
# Evaluation

- Interviewed 27 locals
  - Residents, urban planners, businesses
  - Asked them to draw their mental maps of areas first
  - Then showed them our maps and solicited feedback



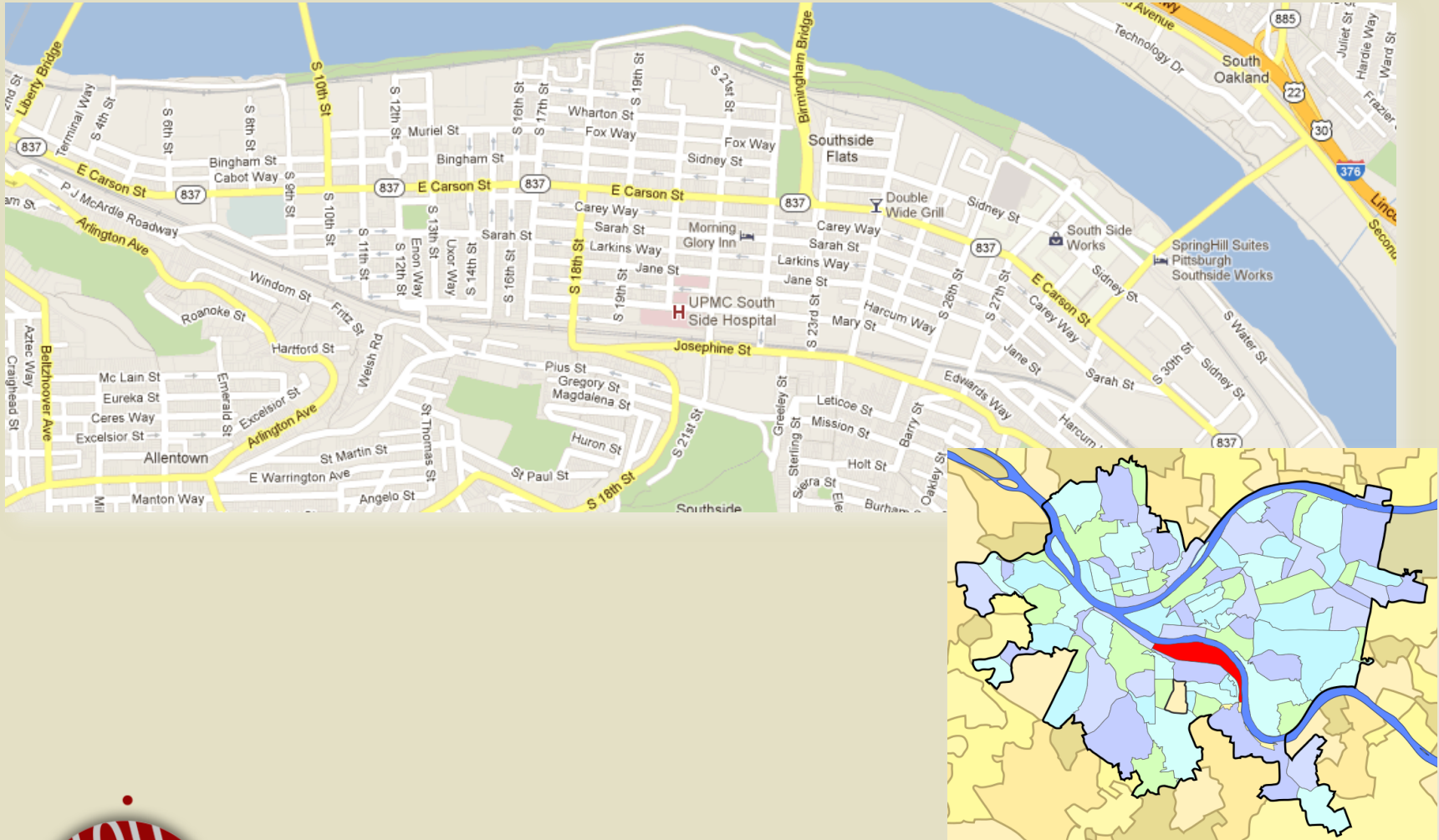


# South Side Pittsburgh

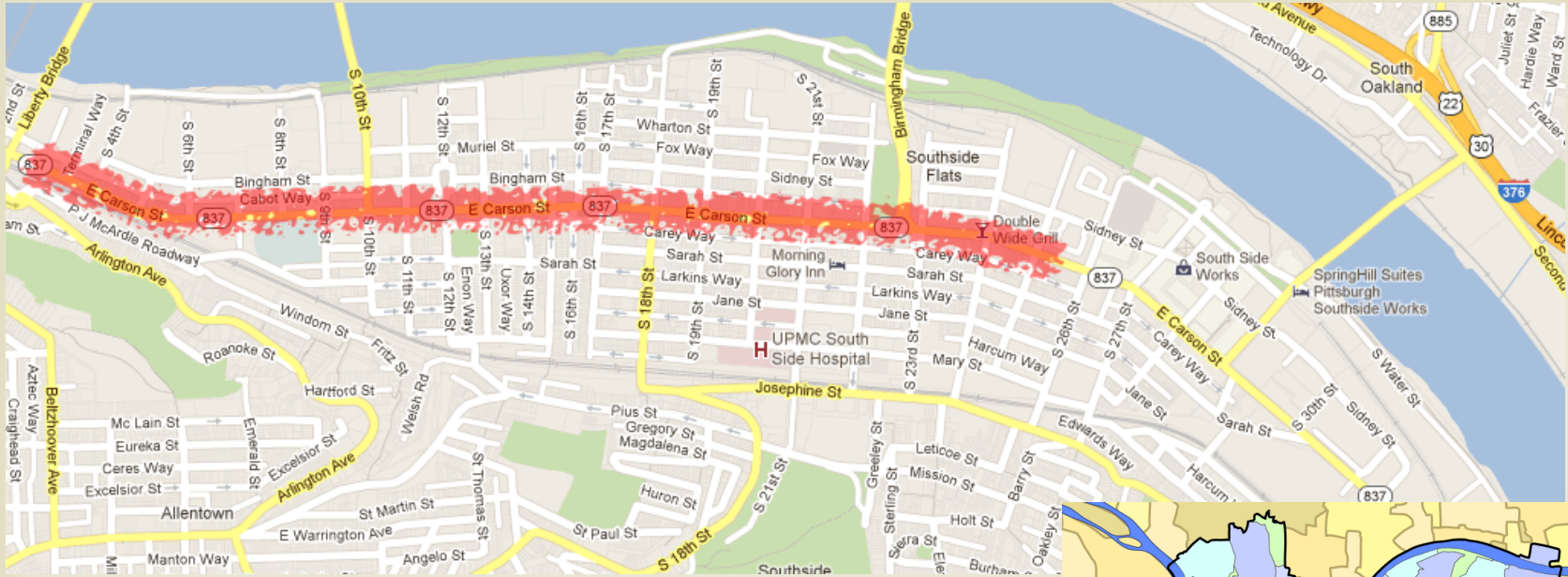




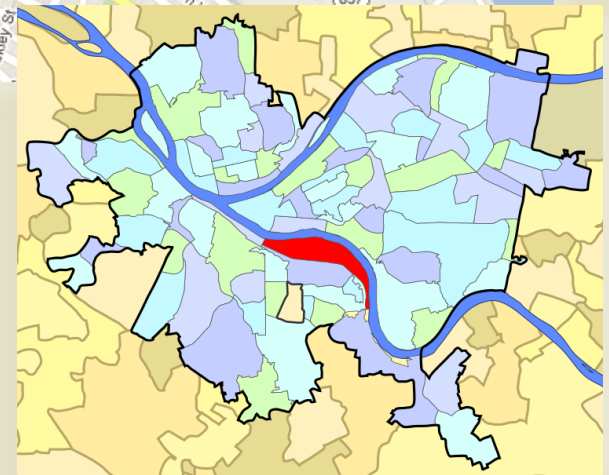
# South Side Pittsburgh



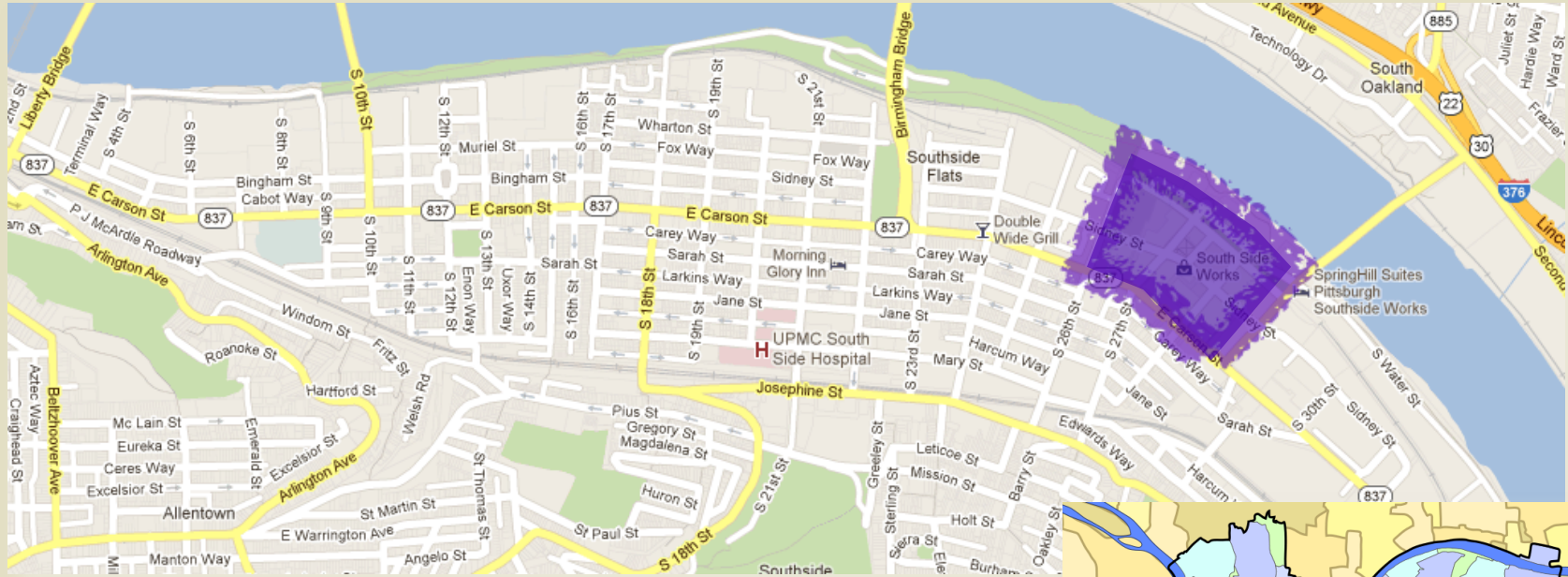
# South Side Pittsburgh



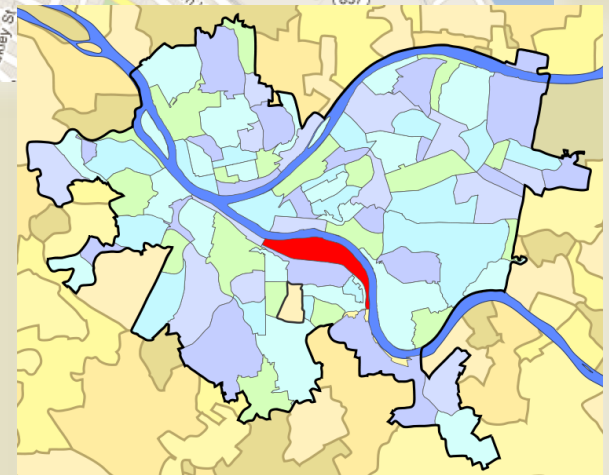
**Carson Street** runs along the length of South Side, and is densely packed with bars, restaurants, tattoo parlors, and clothing and furniture shops. It is the most popular destination for nightlife.



# South Side Pittsburgh

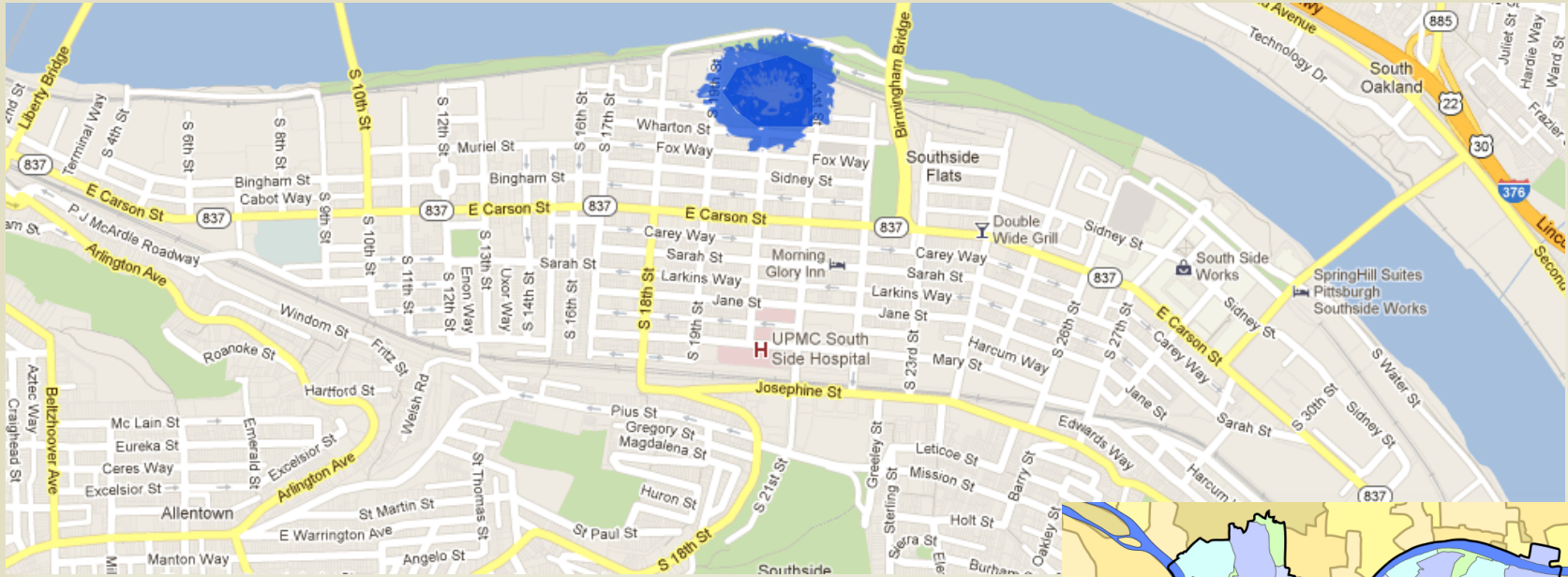


**South Side Works** is a recently built, mixed-use *outdoor shopping mall*, containing nationally branded apparel stores and restaurants, upscale condominiums, and corporate offices.

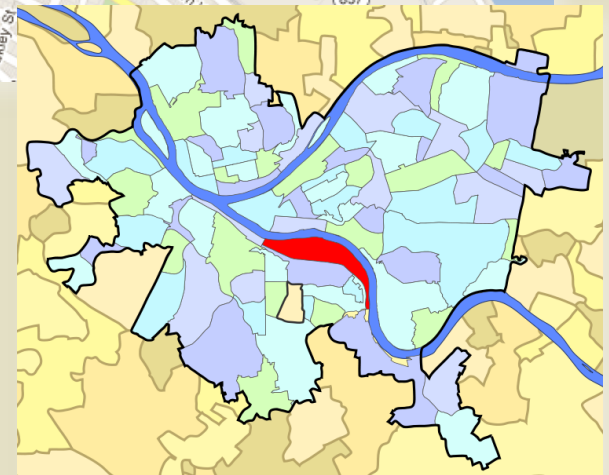




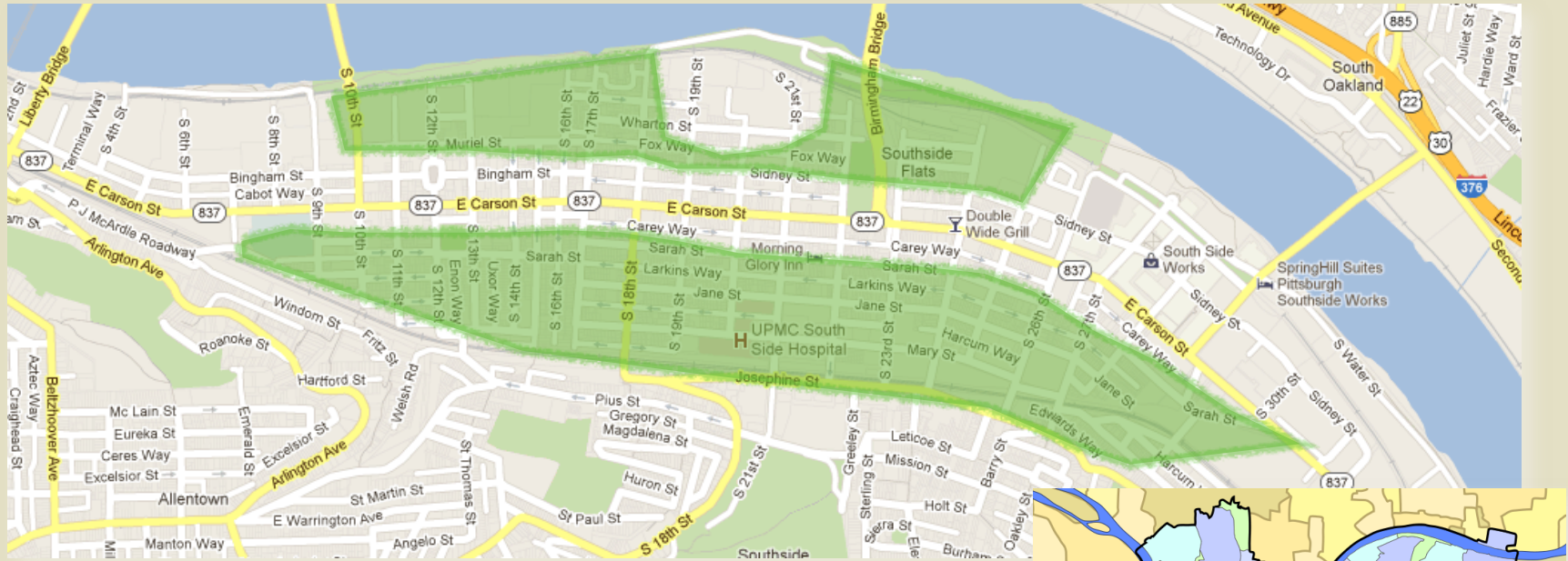
# South Side Pittsburgh



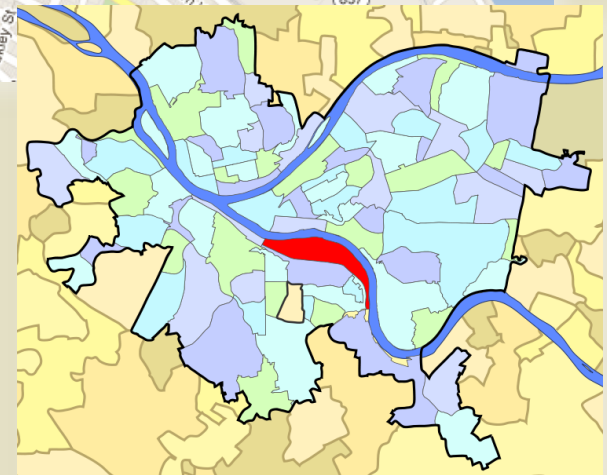
There is an small, somewhat older **strip-mall** that contains the only super market (grocery) in South Side. It also has a liquor store, an auto-parts store, a furniture rental store and other small chain stores.



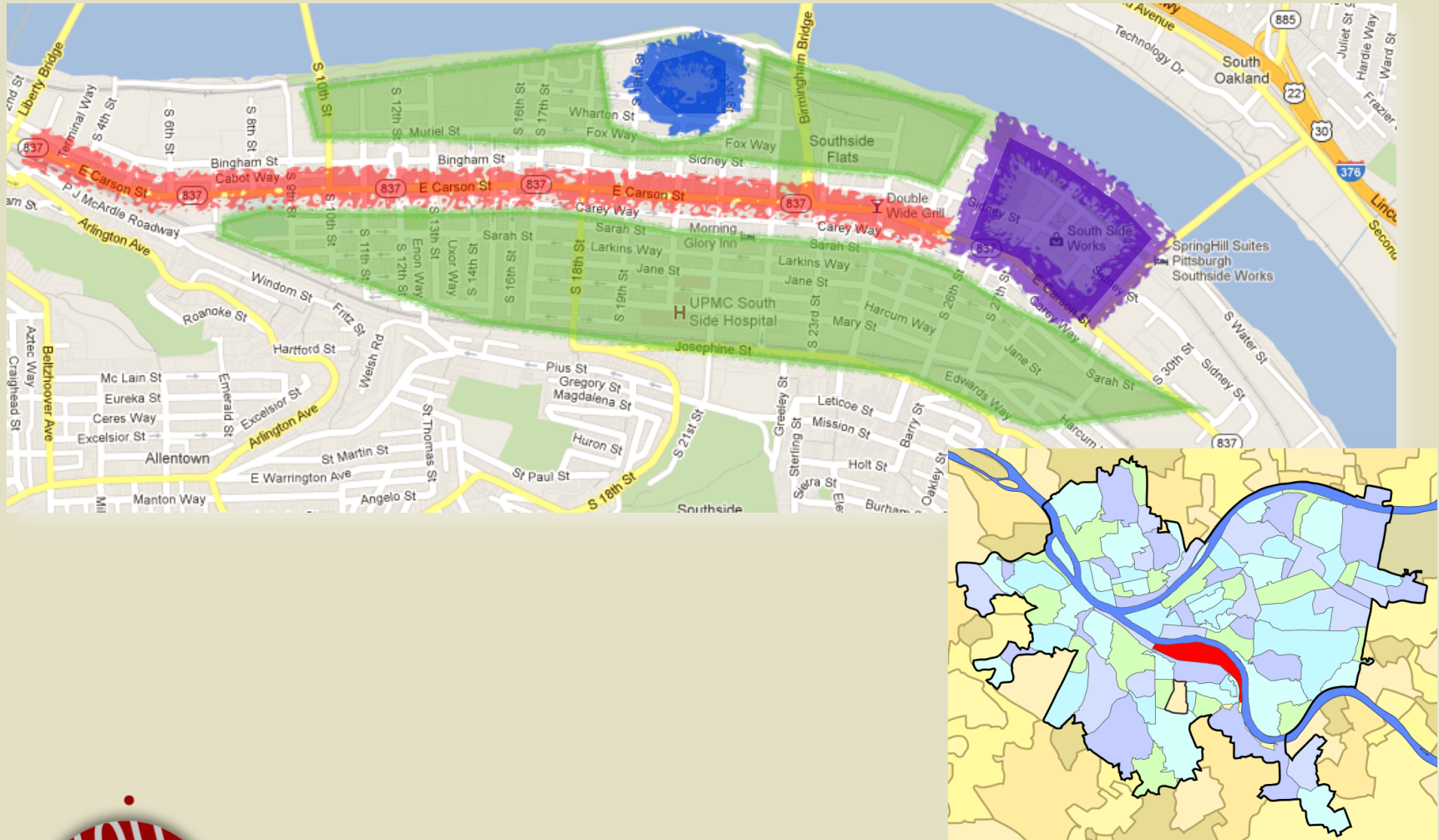
# South Side Pittsburgh



The rest of South Side is predominantly **residential**, consisting of mostly smaller row houses.

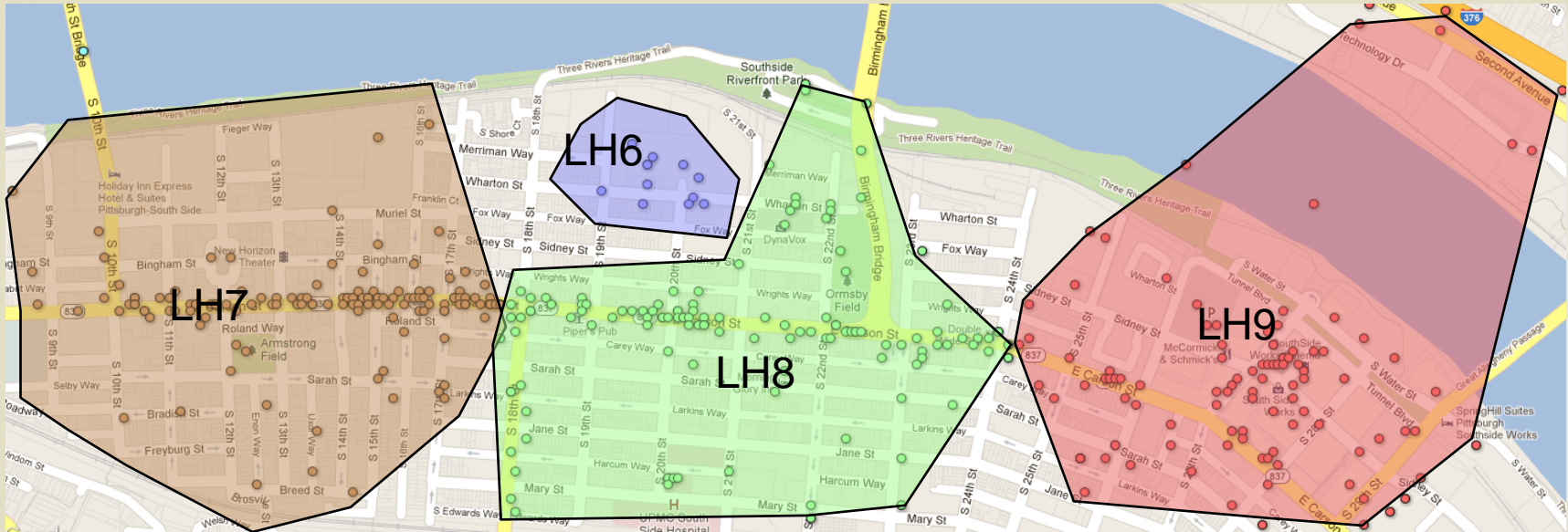


# South Side Pittsburgh





# South Side Pittsburgh

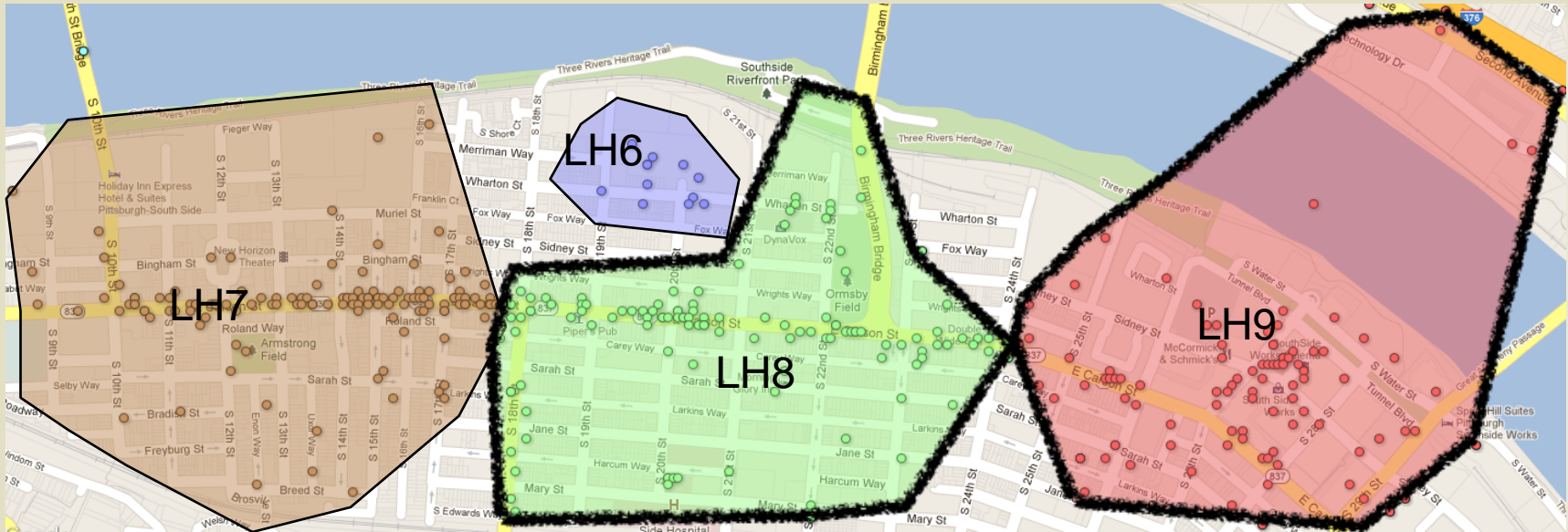


## Livehoods Found in South Side

I'll show evidence in support of the Livehoods clusters in South Side, and will describe the forces that people highlighted.



# South Side Pittsburgh



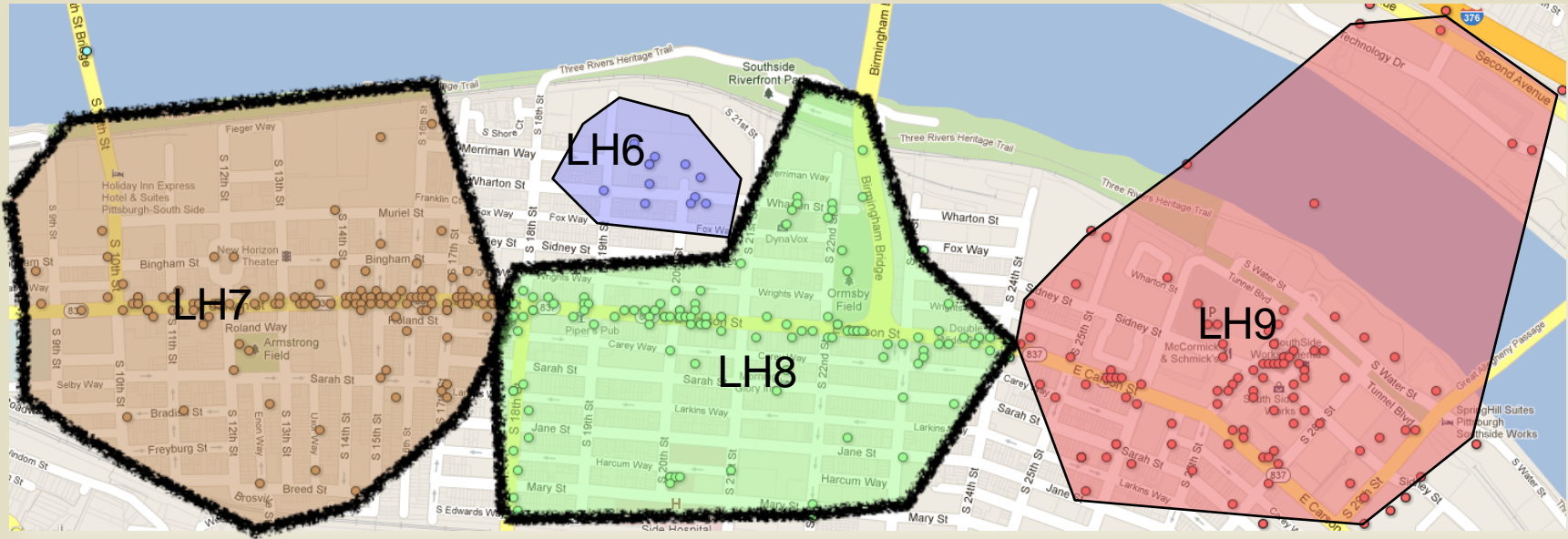
LH8 vs LH9

Demographic  
Differences

“Ha! Yes! See, here is my division! Yay! Thank you algorithm! ... I definitely feel where the South Side Works, and all of that is, is a very different feel.”



# South Side Pittsburgh



LH7 vs LH8

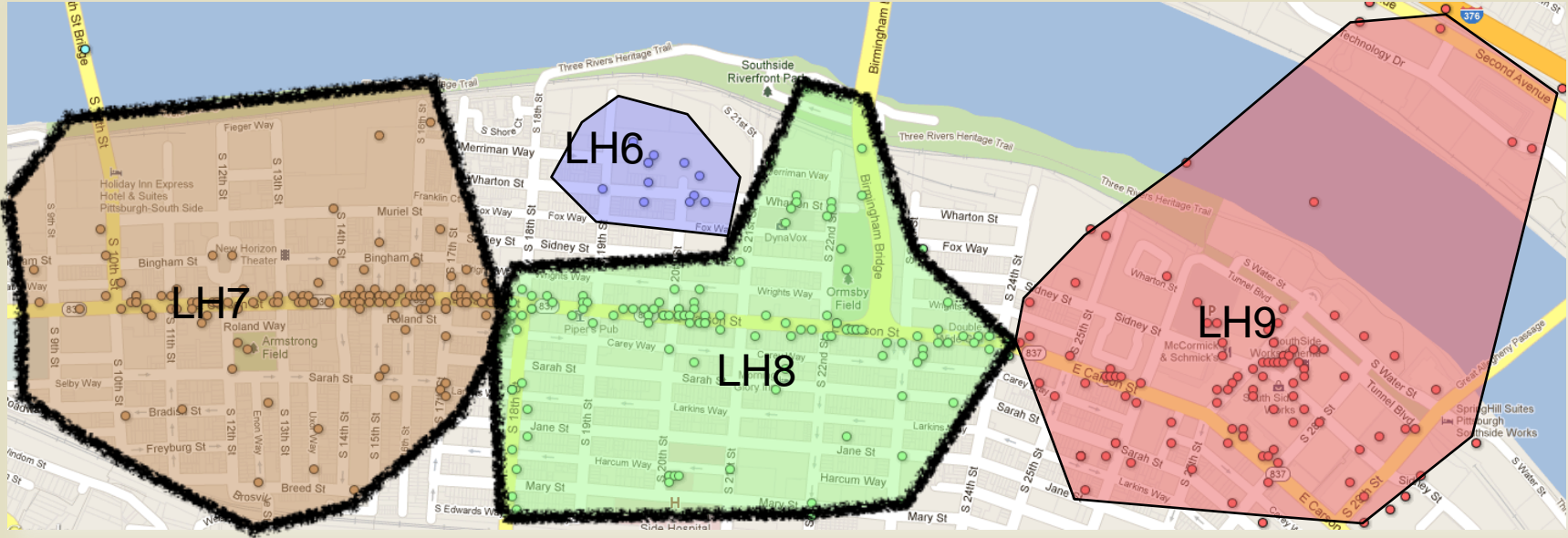
Architecture &  
Urban Design

“from an urban standpoint it is a lot tighter on the western part once you get west of 17th or 18th [LH7].”





# South Side Pittsburgh

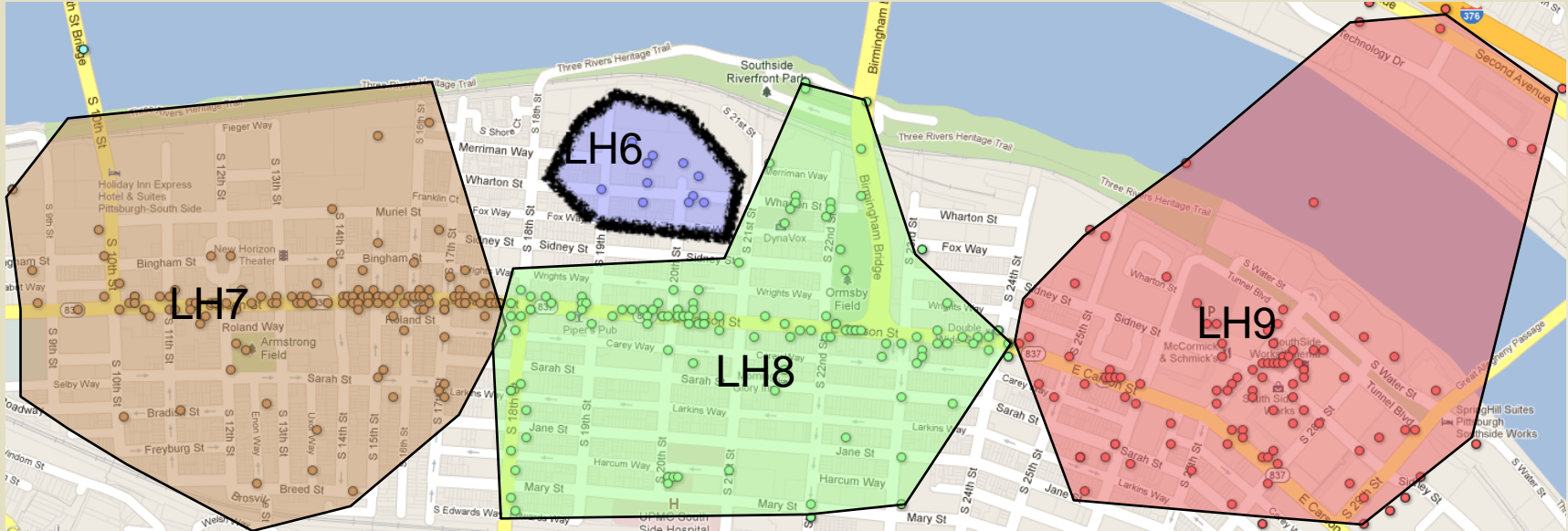


LH7 vs LH8  
Safety

"Whenever I was living down on 15th Street [LH7] I had to worry about drunk people following me home, but on 23rd [LH8] I need to worry about people trying to mug you... so it's different. It's not something I had anticipated, but there is a distinct difference between the two areas of the South Side."



# South Side Pittsburgh



## LH6 Demographic Differences

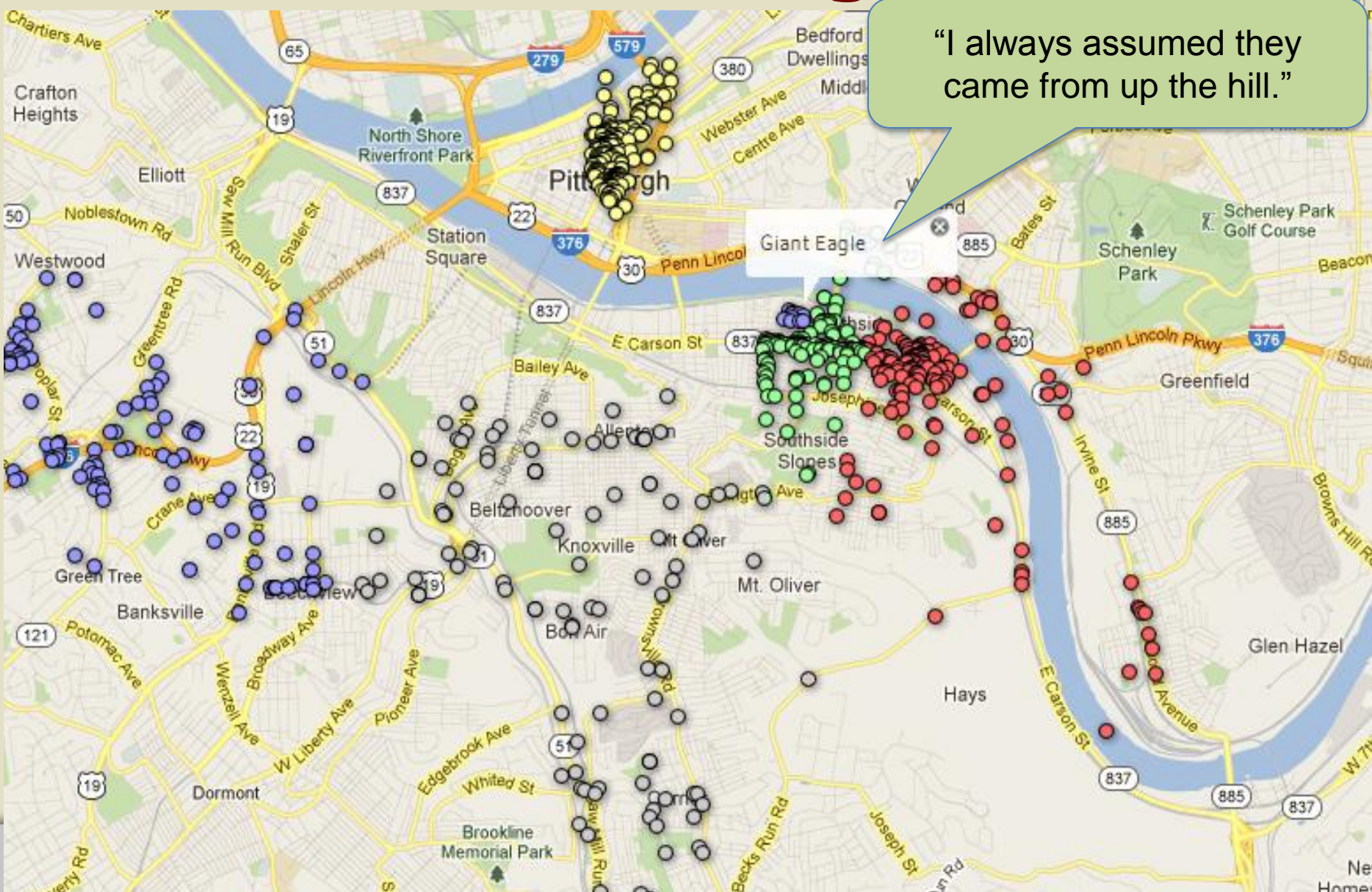
“There is this interesting mix of people there I don’t see walking around the neighborhood. I think they are coming to the Giant Eagle [grocery store] from lower income neighborhoods... **I always assumed they came from up the hill.**”





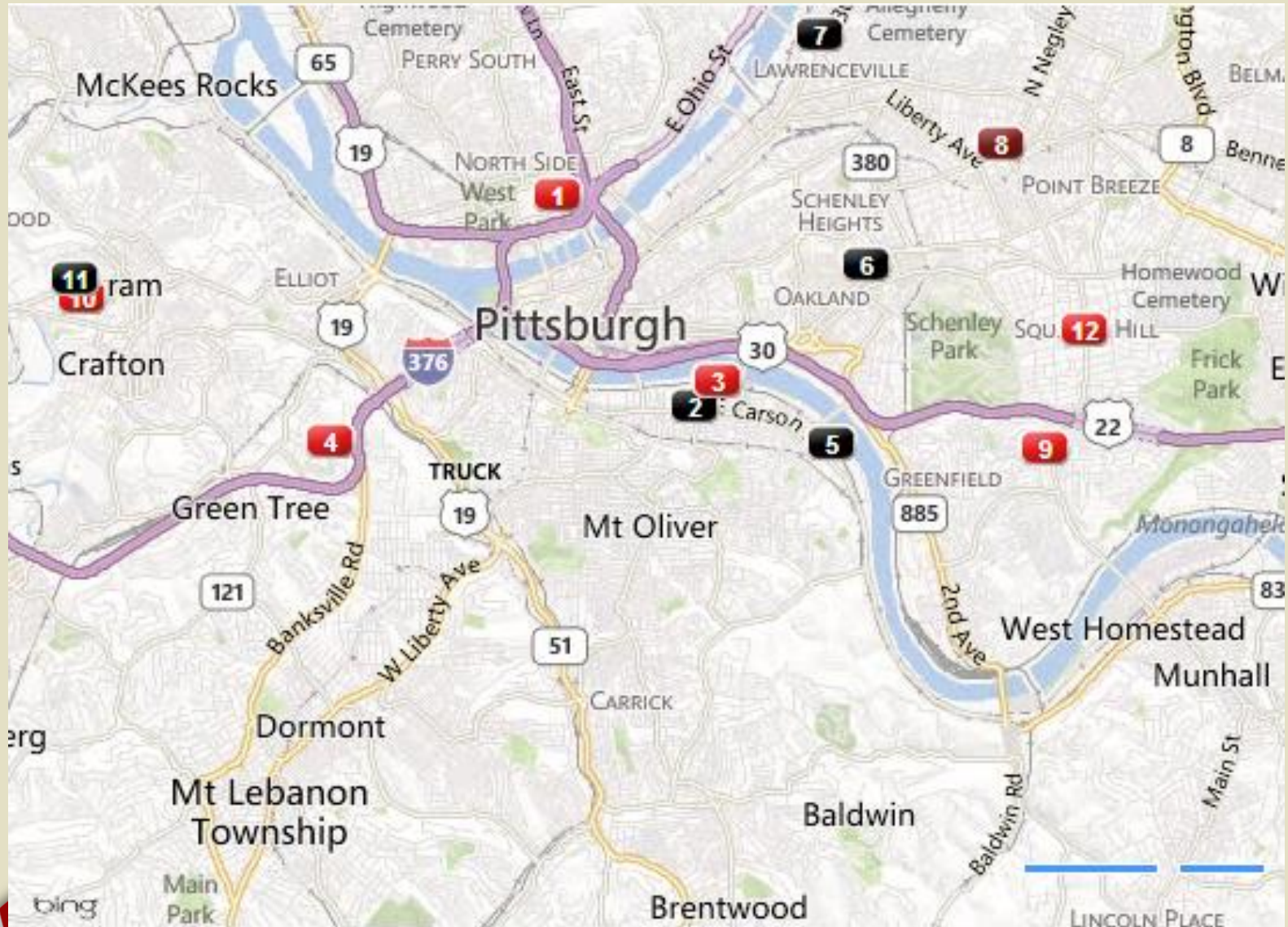
# South Side Pittsburgh

"I always assumed they came from up the hill."



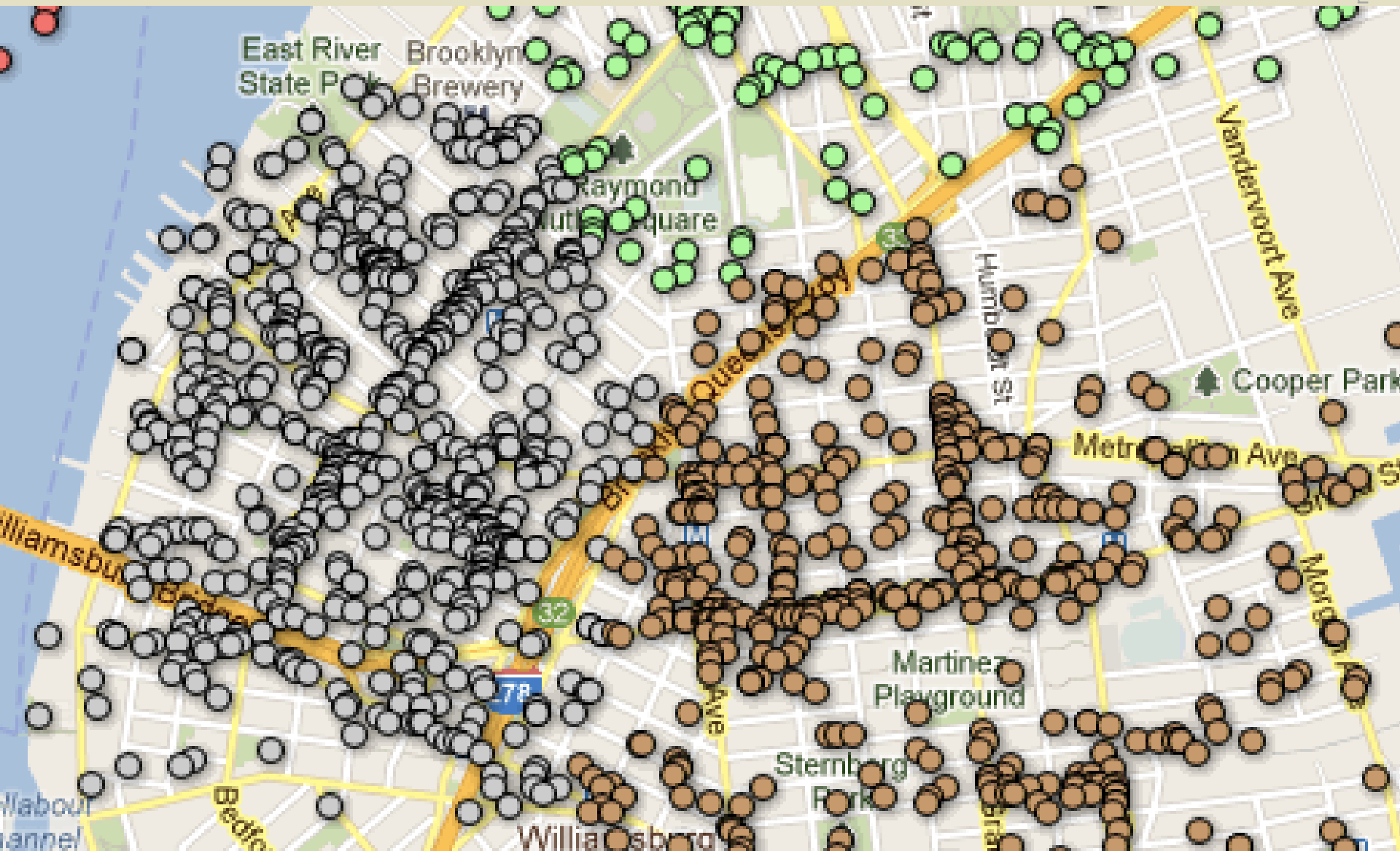


# South Side Pittsburgh

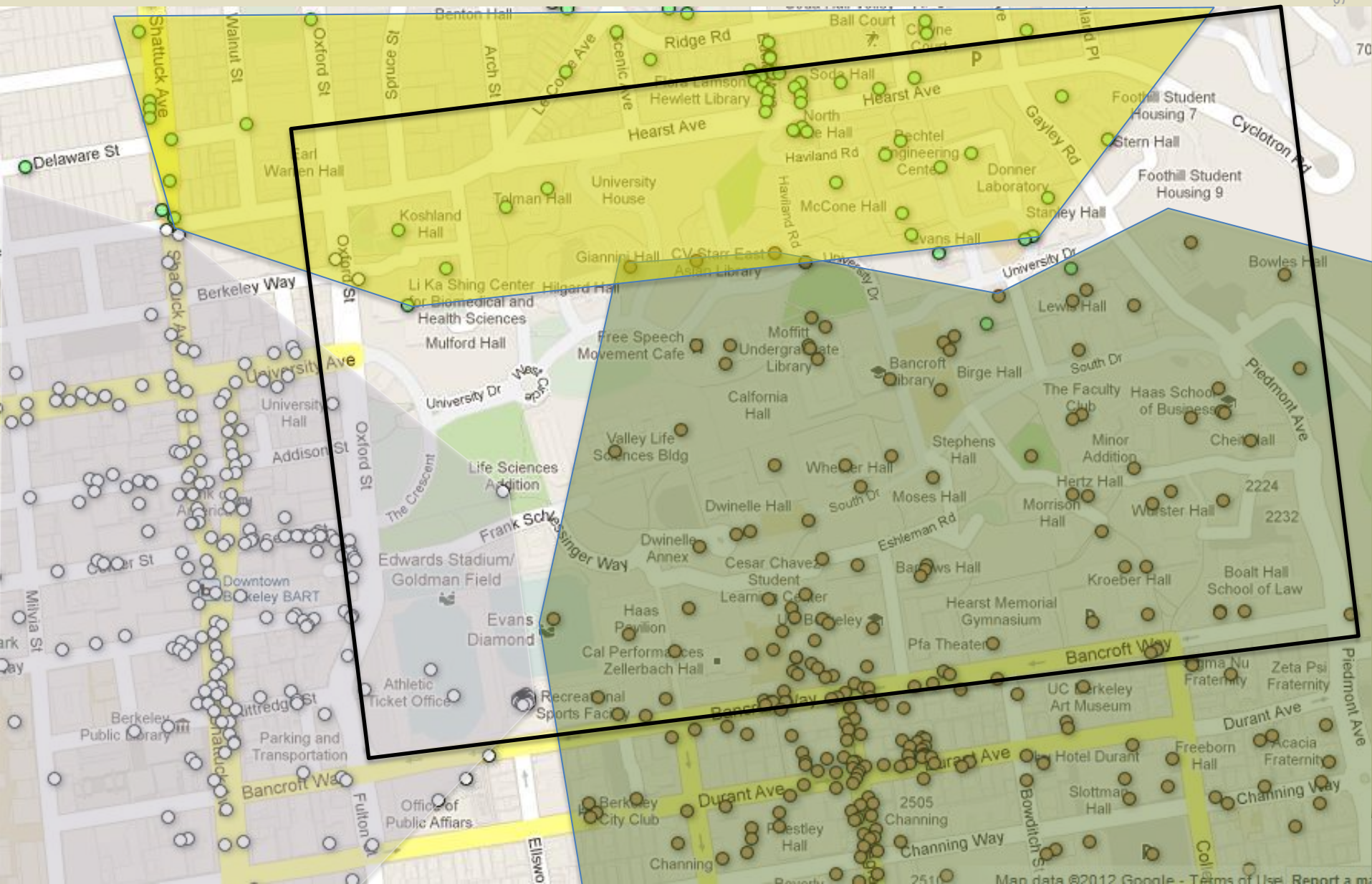




# Brooklyn Queens Expressway

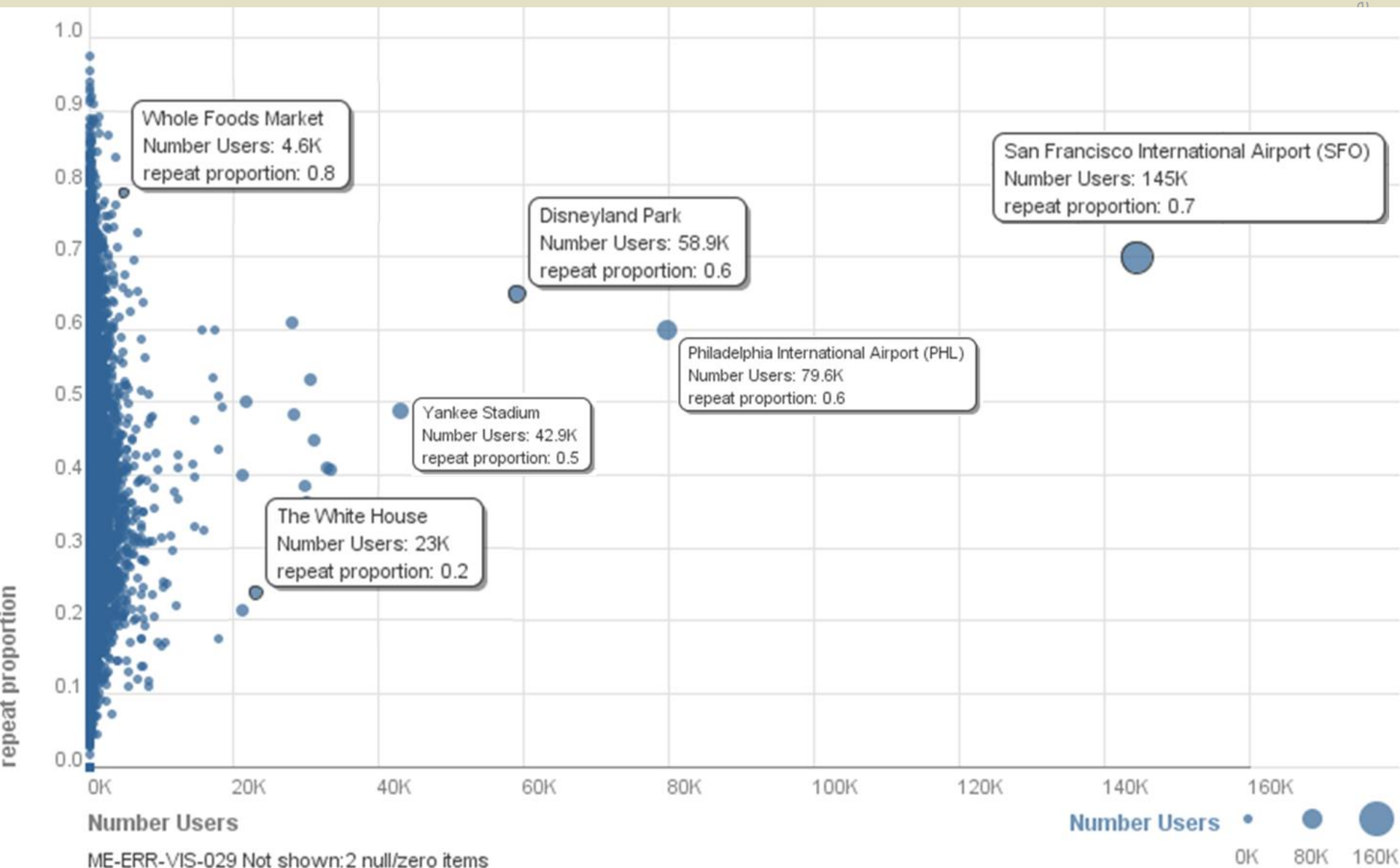


# Bezerkeley, CA

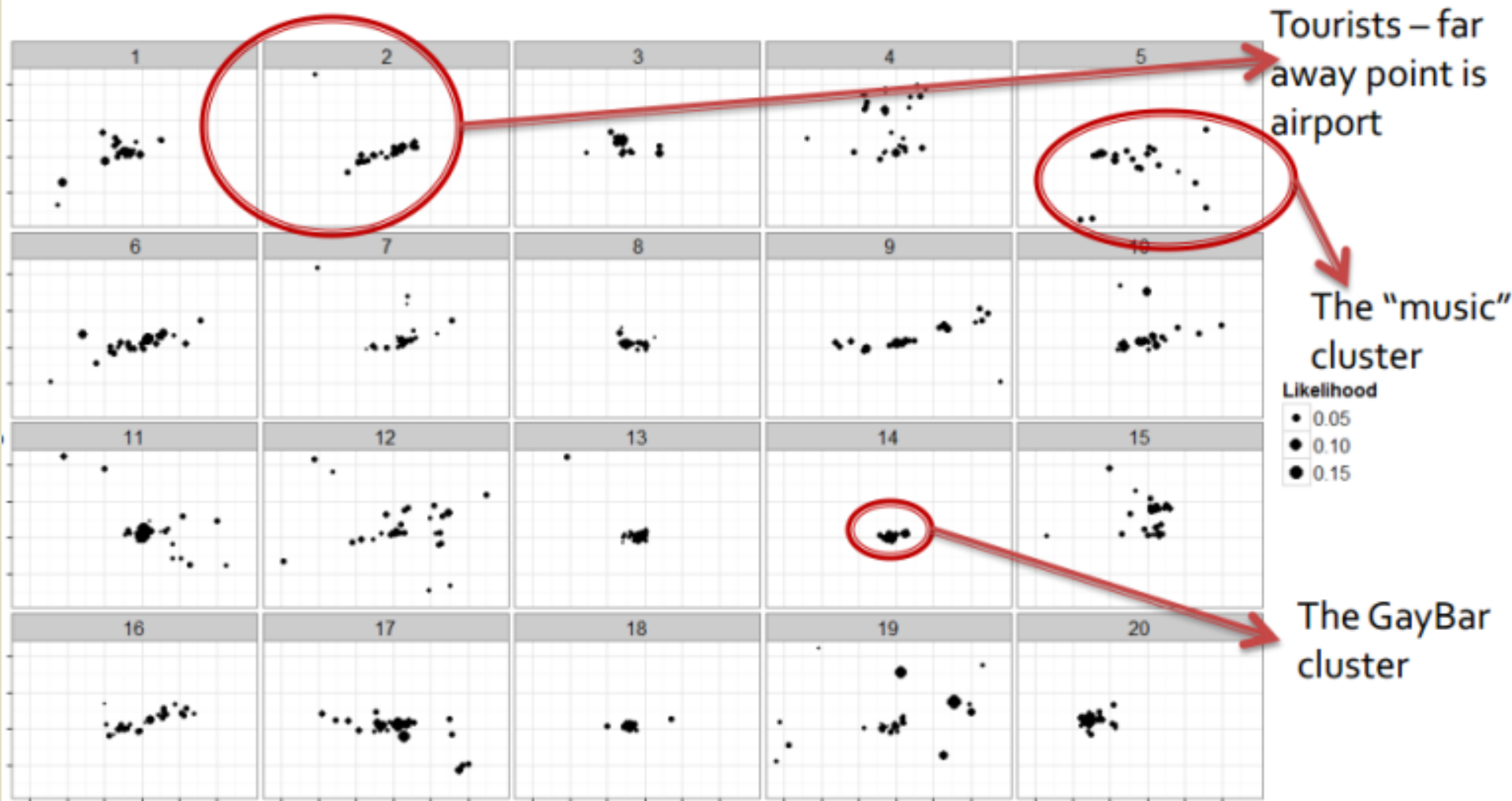




# Other Potential Urban Analytics



# Topic Modeling (LDA)





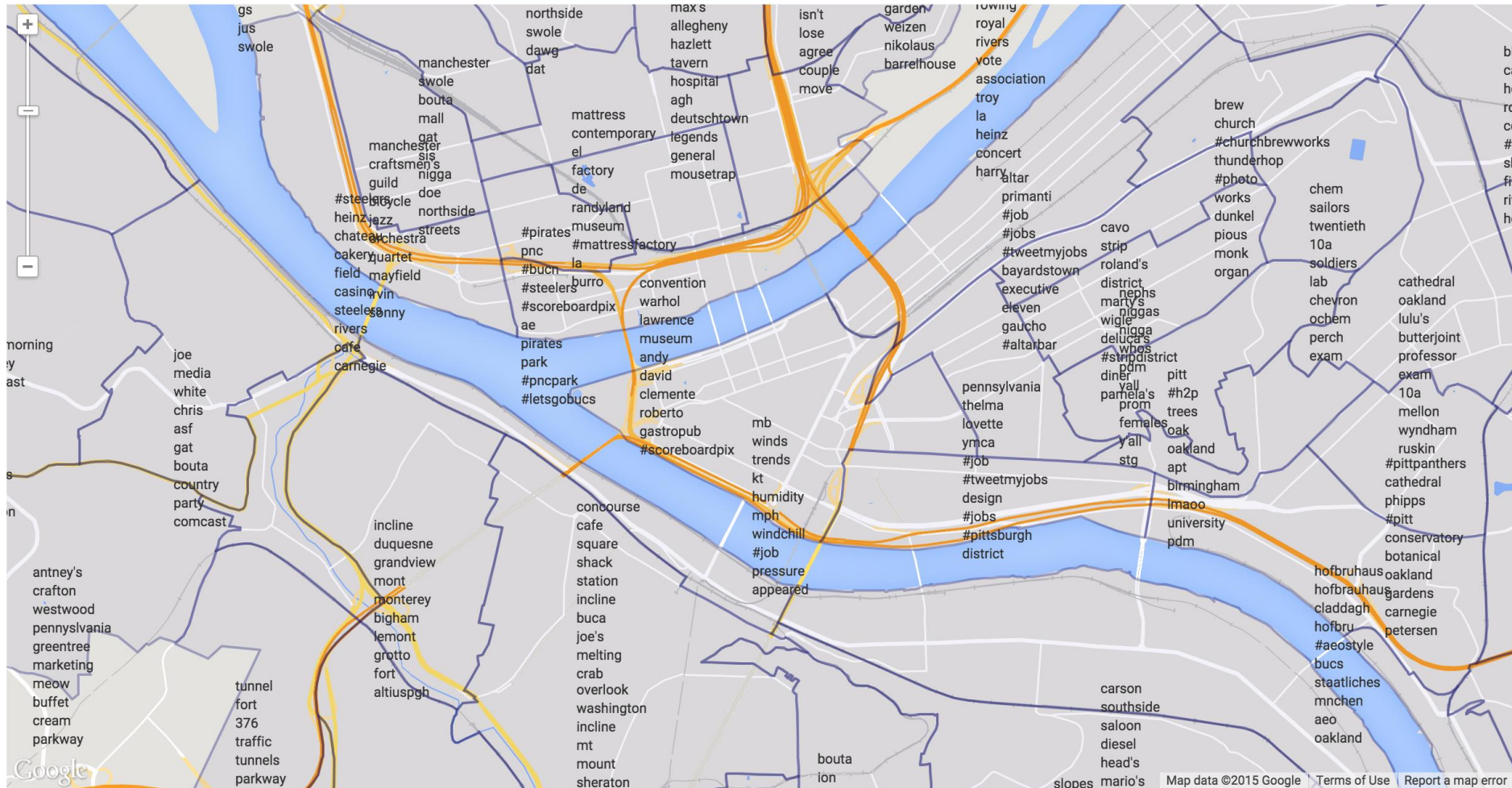
# Analysis of Geotagged Tweets in Pittsburgh

Neighborhood Names

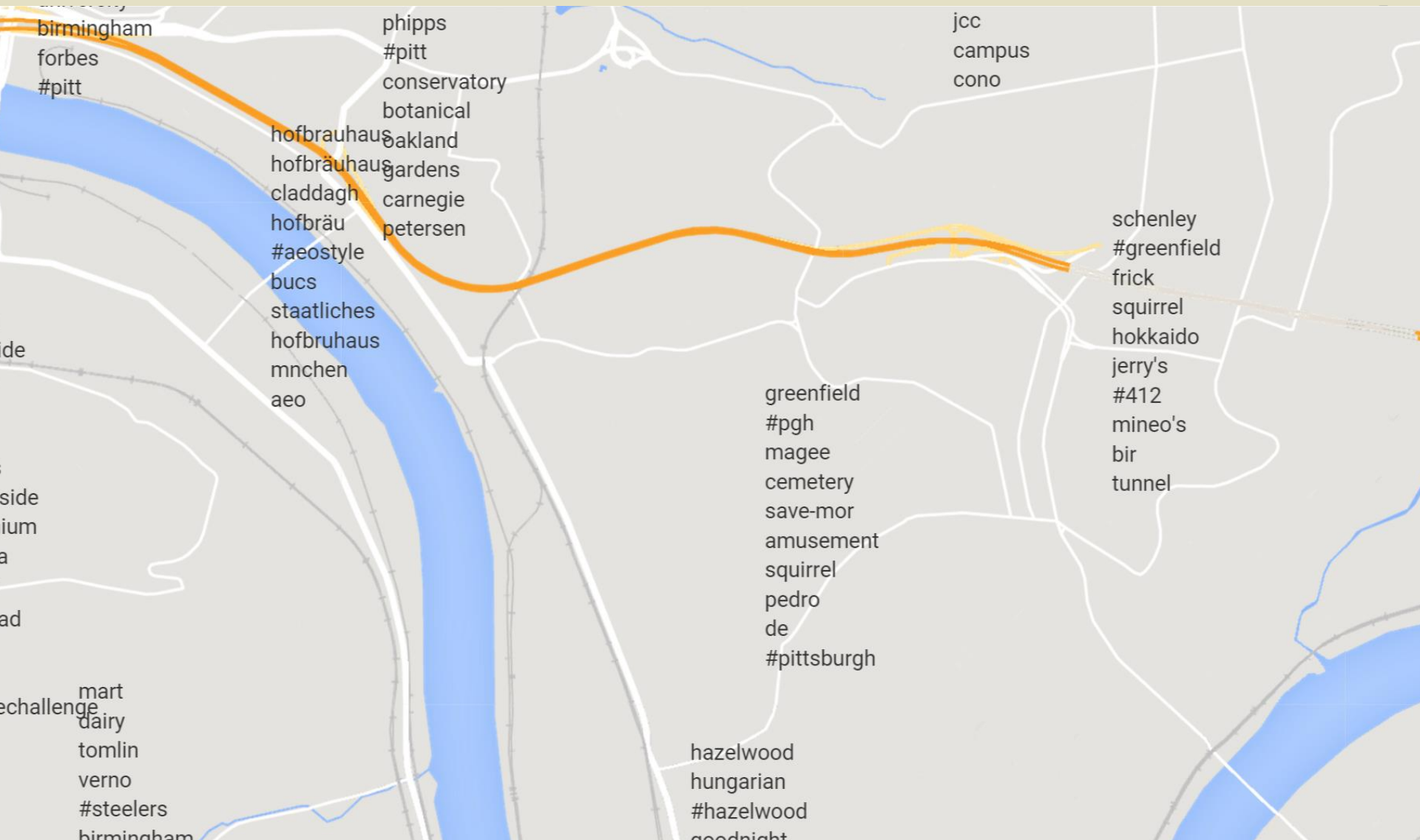
Top 3 Emojis per Neighborhood

Top 10 Words per Neighborhood

Clear Map

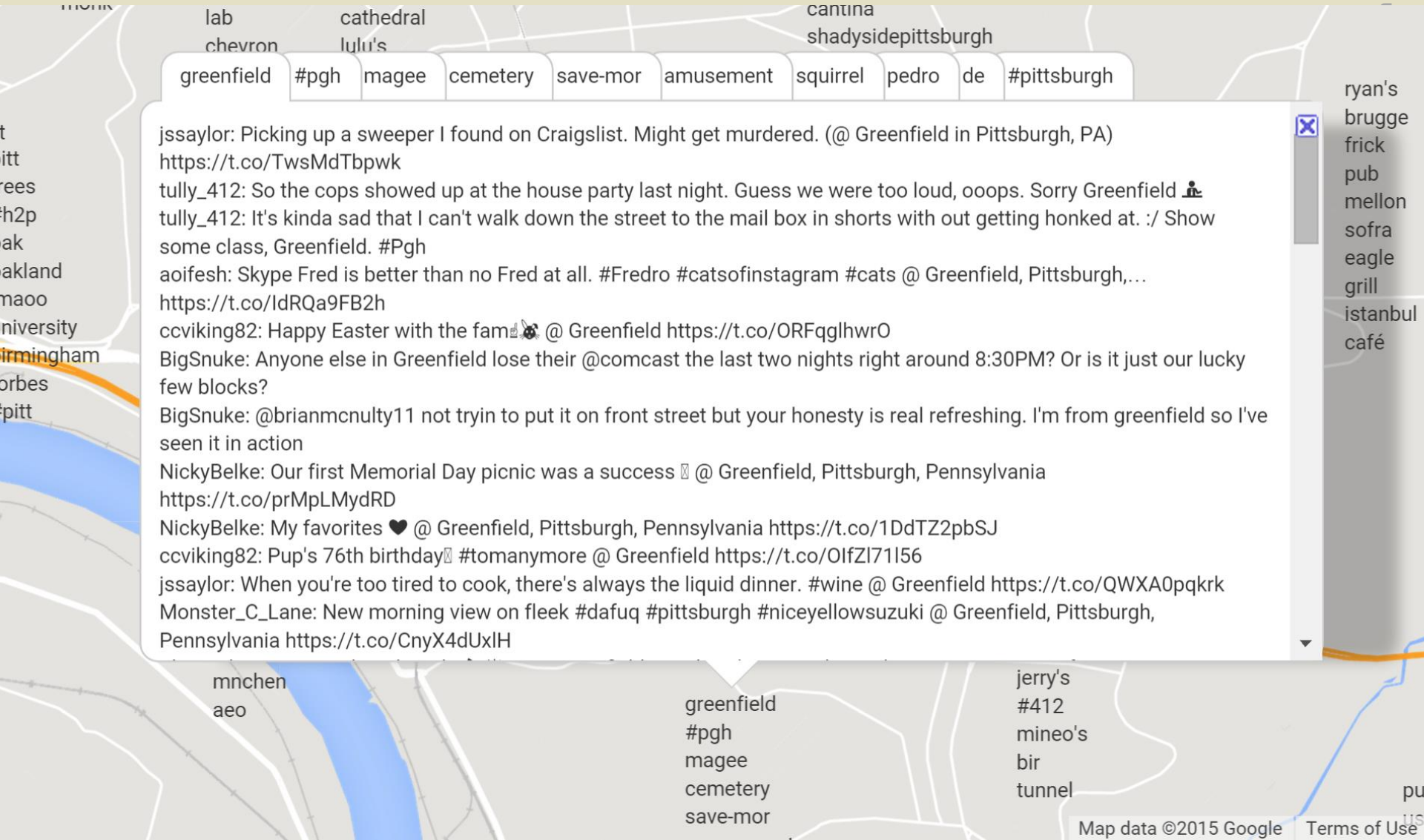


# Analysis of Geotagged Tweets in Pittsburgh





# Analysis of Geotagged Tweets in Pittsburgh



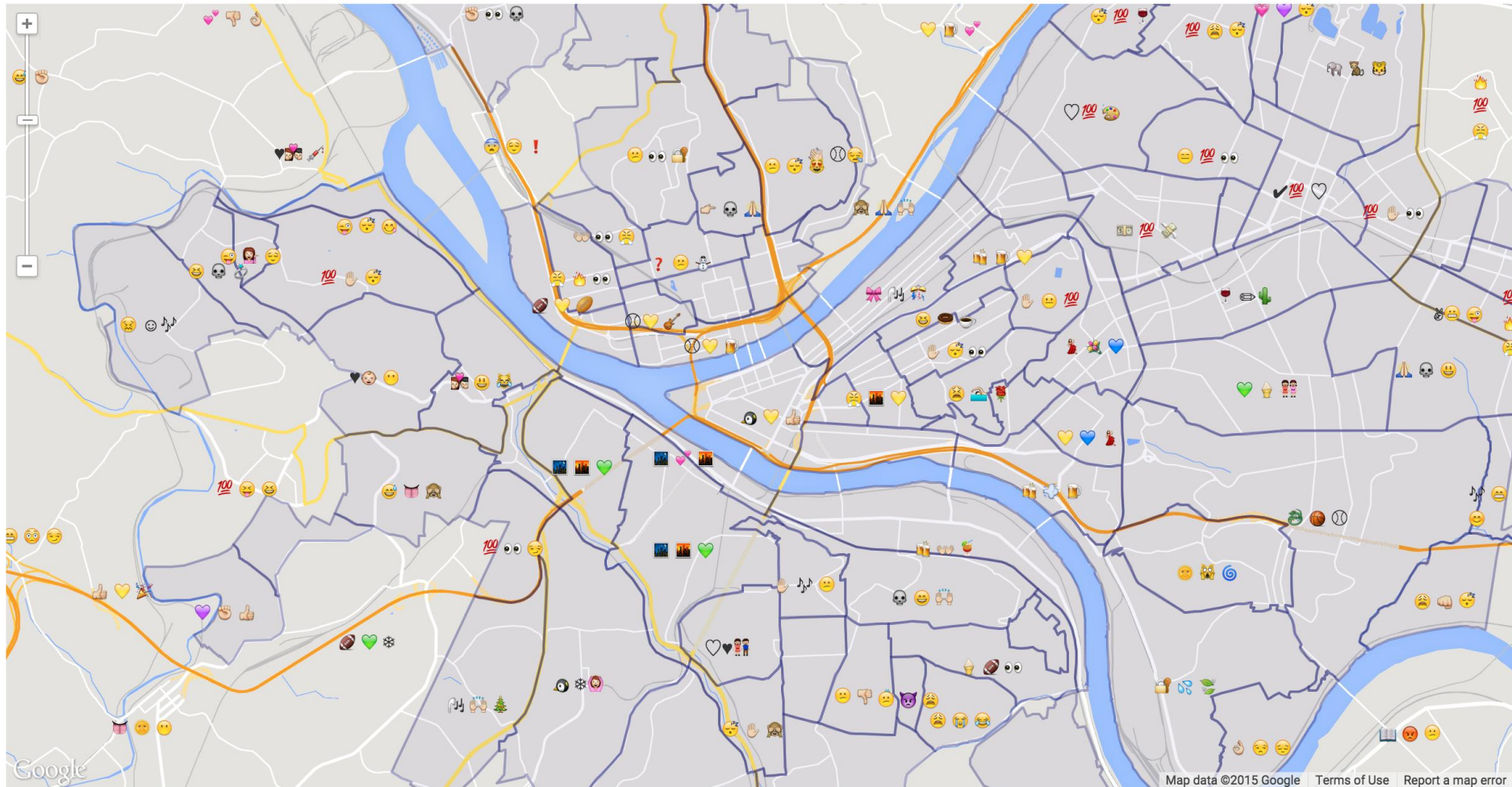
# Analysis of Geotagged Tweets in Pittsburgh

Neighborhood Names

Top 3 Emojis per Neighborhood

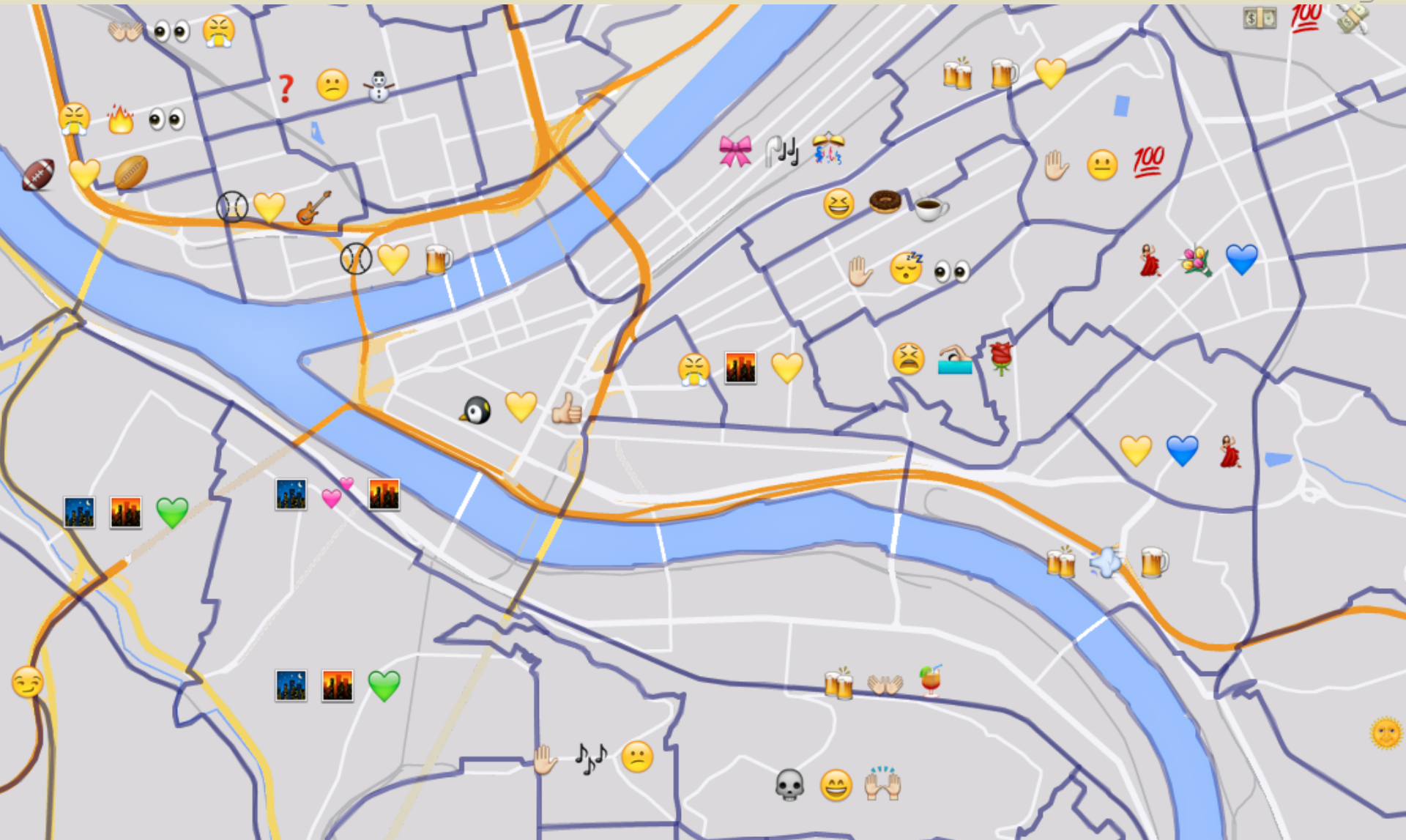
Top 10 Words per Neighborhood

Clear Map



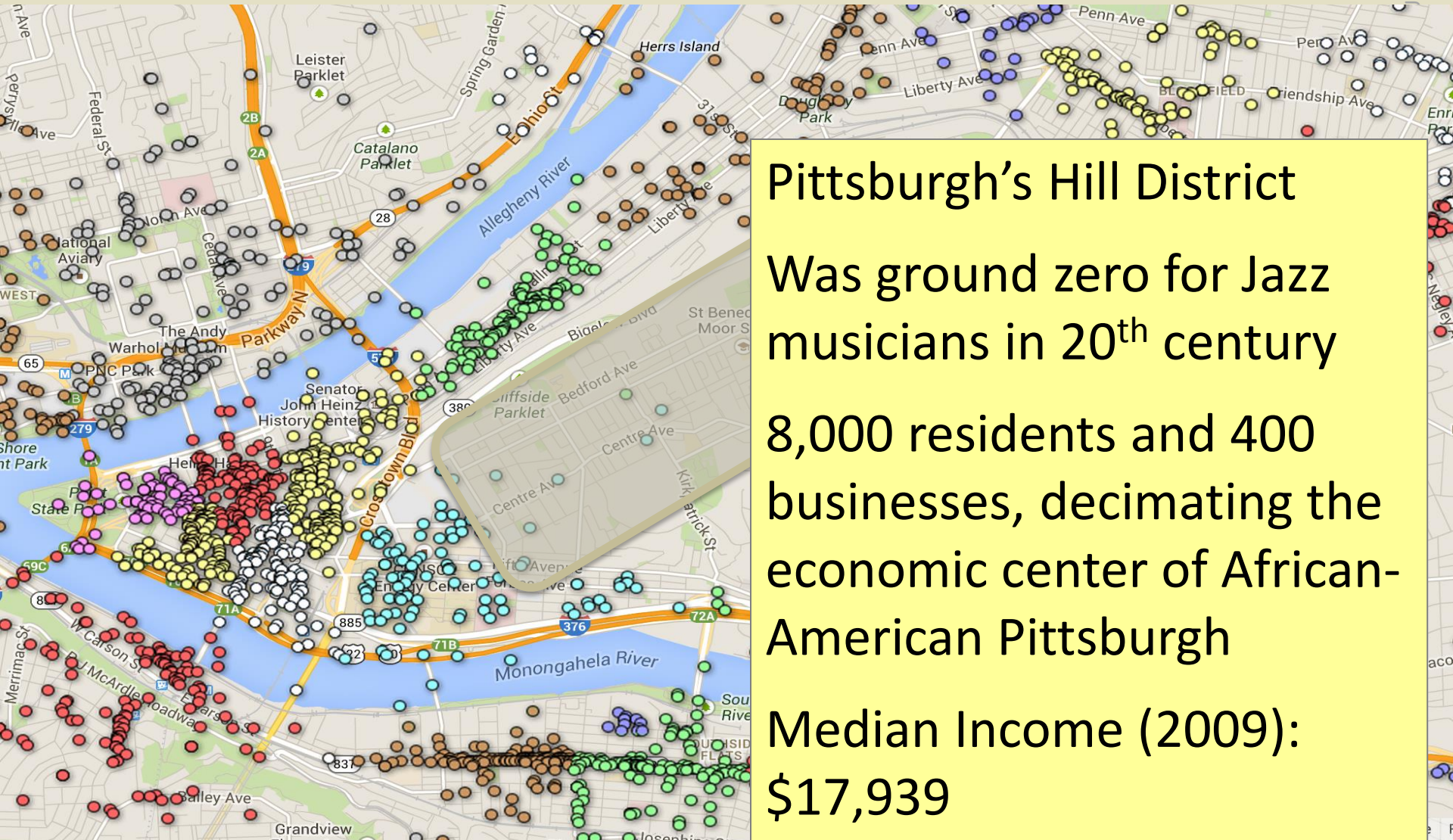


## University : 50



# Reflections on Urban Analytics

## *Potential Biases in the Data?*





# Reflections on Urban Analytics

## *Potential Biases in the Data?*

- Socioeconomic bias
  - Little foursquare data in lower socioeconomic areas
  - (Less of a problem with Twitter though)
- Urban bias
  - Social media more active per capita in cities
- Age and gender bias
  - Most young, male, technology-savvy
- Is this a problem that will solve itself with time?
  - Or, can we address this in our models?
  - Or, use multiple sources of social media data



# Reflections on Urban Analytics

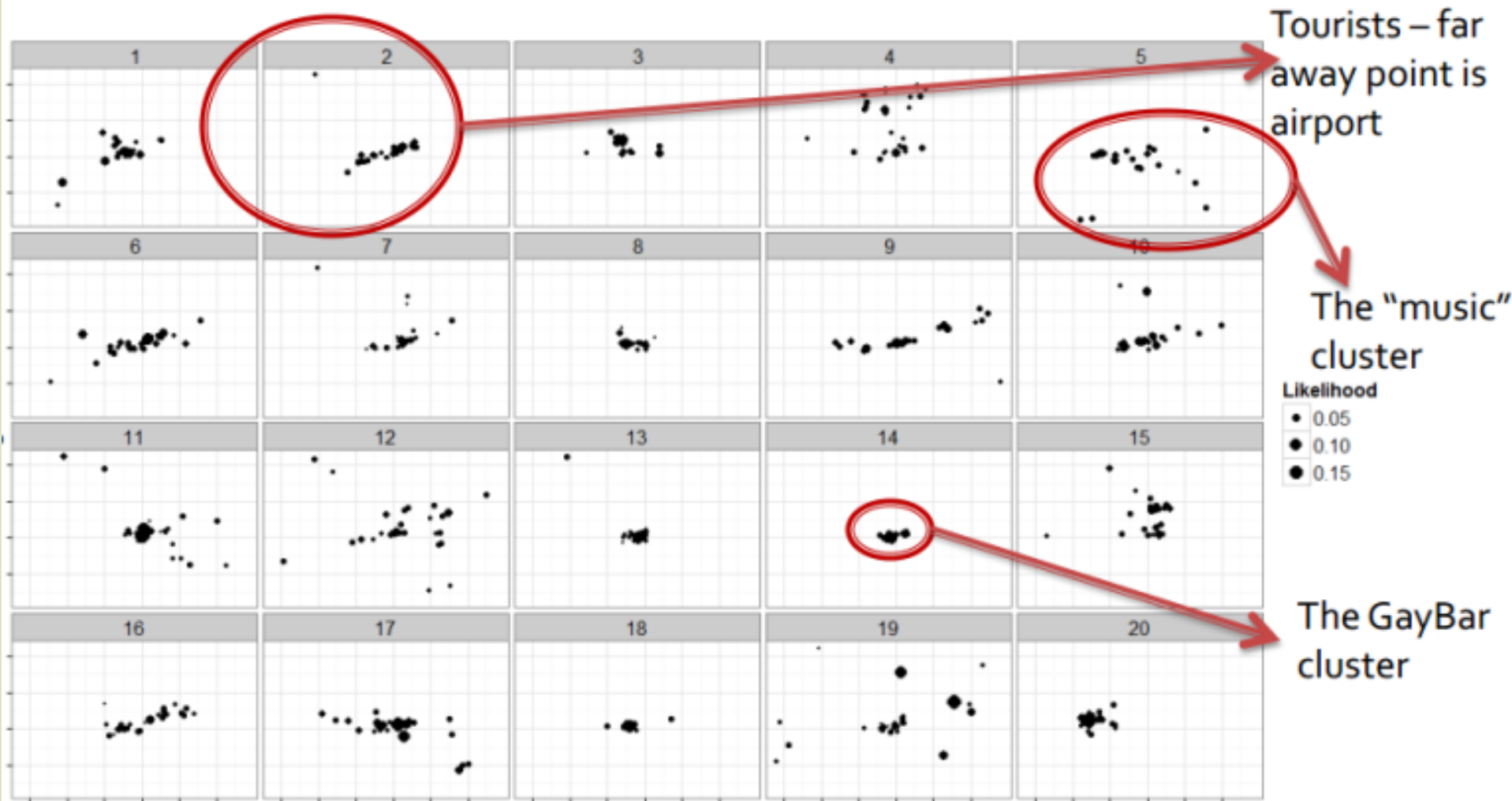
## *Privacy Concerns?*

- Publicly visible data without requiring logins
  - No IRB issues
- Removed venues labeled as “home”
  - We only received one request to remove a venue from Livehoods (wasn’t labeled as a home)
- We only show data about geographic areas vs individuals
  - Can’t identify behaviors of specific individuals
- But still many other questions





# How Much Can Be Inferred?



# How Much Can Be Inferred?

- Very likely much more can be inferred using rich data like this
  - Demographics, socioeconomic, friends
  - Physical and mental health (depression)
  - How “risky” you are (bars, clinics, etc)
- Unclear how far inferencing can go
  - Also, not much can stop advertisers, NSA, startups,
  - Even if an individual hides behaviors, can infer a lot based on what other similar people are doing





# How Much Can Be Inferred?

The screenshot shows the homepage of the website "f i @ s t m x ñ d @ ¥". The header features the site's name in two rows: "f i ® s t" on a red background and "m x ñ d @ ¥" on a black background. Below this is the text "PEER-REVIEWED JOURNAL ON THE INTERNET" followed by a dashed line. A navigation bar contains links: ABOUT, LOGIN, REGISTER, SEARCH, CURRENT, ARCHIVES, ANNOUNCEMENTS, and SUBMISSIONS. Below the navigation bar, it says "> Volume 14, Number 10 - 5 October 2009 > Jernigan". At the bottom of the screenshot, the site's name and "PEER-REVIEWED JOURNAL ON THE INTERNET" are repeated.

Built a logistic regression to predict sexuality based on what your friends on Facebook disclosed

Gaydar: Facebook friendships expose sexual orientation

by Carter Jernigan and  
Behram F.T. Mistree



[sign in](#) [new guest?](#) [my account](#) [REDCard](#)

search



[women](#) [men](#) [baby](#) [kids](#) [home](#) [patio](#) [furniture](#) [electronics](#) [entertainment](#) [toys](#) [health & beauty](#) [clearance](#) [more](#)

REDCard



**SAVE 5% + GET FREE SHIPPING.**  
TODAY & EVERYDAY - Apply today

[find a store](#)

[Weekly Ad](#)

[GiftCards](#)

[registries](#)

[TargetLists](#)



## baby

**save 20%** when you  
spend \$75

[girls' clothing](#) | [boys' clothing](#) | [offer details](#)

**save 10%** when you  
buy 2 swim diapers  
or baby sunscreen.

[swim diaper deals](#) | [sunscreen deals](#) | [offer details](#)

**free B-safe** infant car seat,  
with B-ready  
stroller purchase.

[Britax deals](#)

### clothing & shoes

[baby & toddler boys'](#)

[clothing](#)

[baby & toddler](#)

[clothing](#)

[boys' shoes](#)

[girls' shoes](#)

### baby gear

[activity gear](#)

[bouncers &](#)

[car seats](#)

[infant carriers](#)

[strollers](#)

[swings](#)

### baby basics

[baby bath](#)

[diapering](#)

[feeding](#)

“[An analyst at Target] was able to identify about 25 products that... allowed him to assign each shopper a ‘pregnancy prediction’ score. [H]e could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy.” (NYTimes)

or Baby,  
rally.

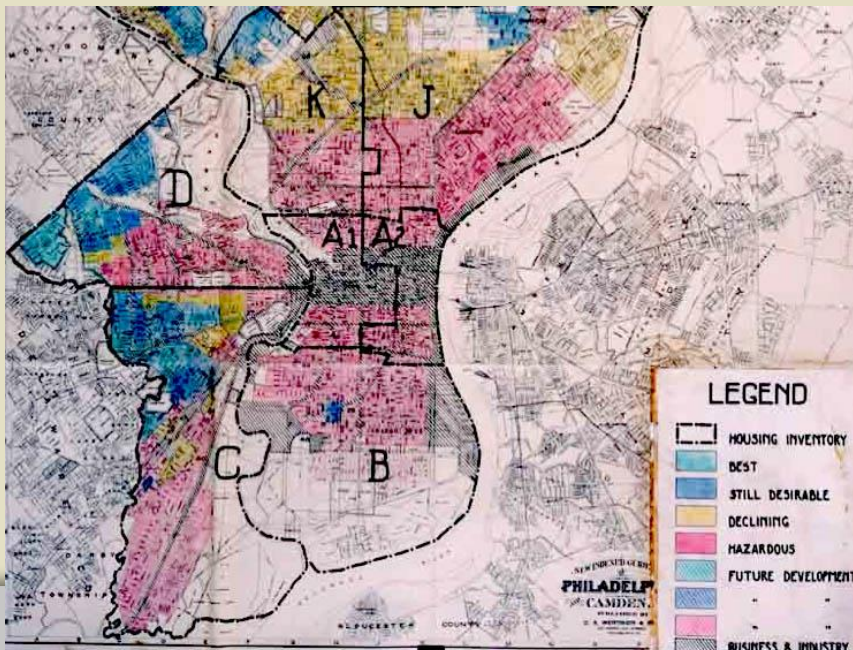
eneration  
aby naturally.

eneration



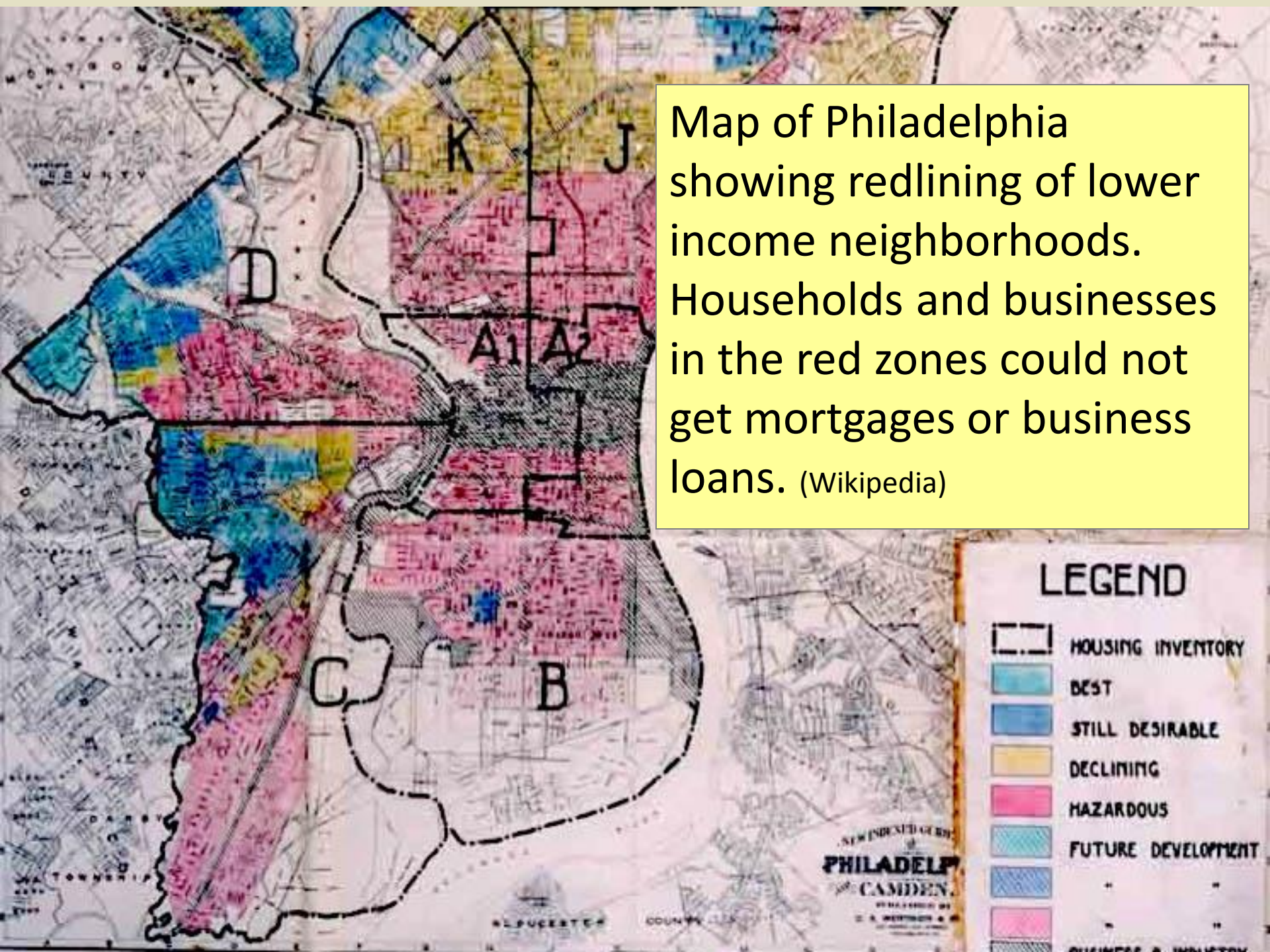
# A New Kind of Redlining?

- “denying, or charging more for, services such as banking, insurance, access to health care, ... supermarkets, or denying jobs ... against a particular group of people” (Wikipedia)





Map of Philadelphia showing redlining of lower income neighborhoods. Households and businesses in the red zones could not get mortgages or business loans. (Wikipedia)





# 'GMA' Gets Answers: Some Credit Card Companies Financially Profiling Customers

Jan. 28, 2009

By CHRIS CUOMO, JAY SHAYLOR, MARY McGUIRT and CH

 Like  share  Tweet  +1



Johnson says his jaw dropped when he read one of the reasons American Express gave for lowering his credit limit:

"Other customers who have used their card at establishments where you recently shopped have a poor repayment history with American Express."

# Moving Forward

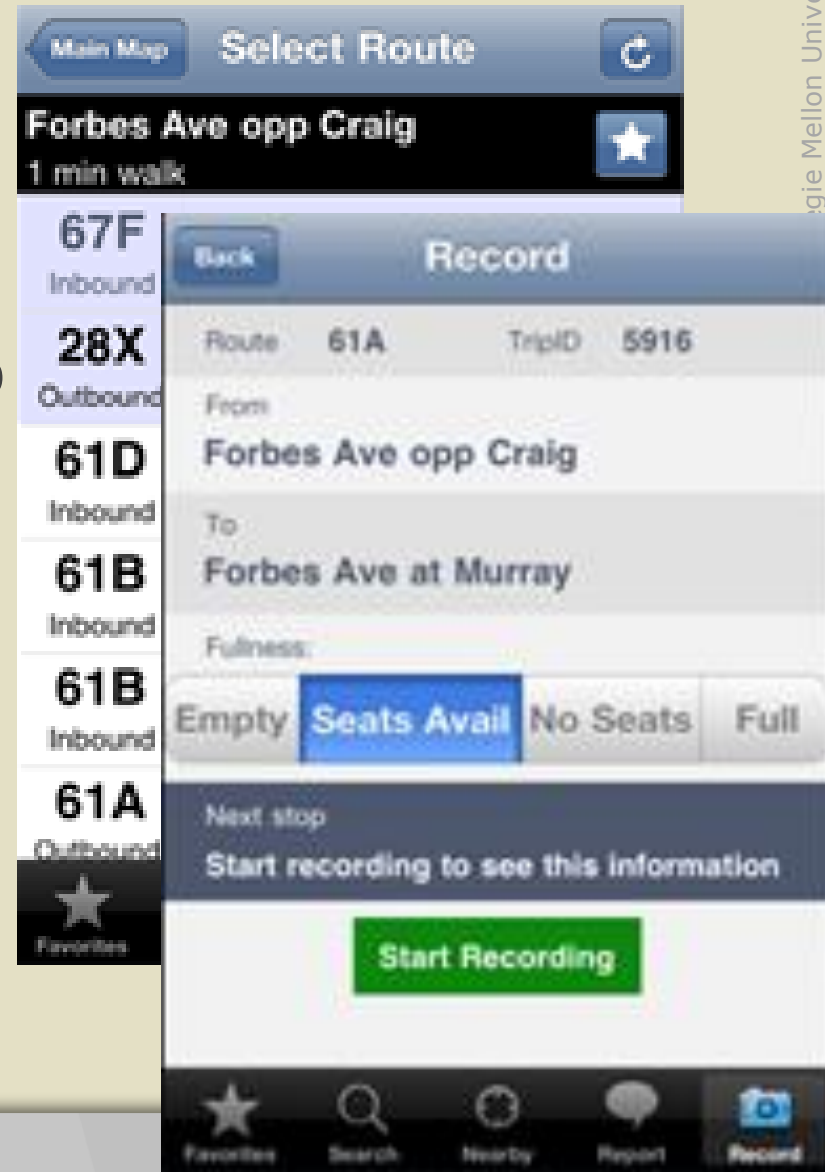
- At early stage, but lots of potential for understanding wide range of behaviors for cities
  - Business analytics, use of city resources (parks), how neighborhoods change over time, health care, location efficiency, and much more
- Many open questions too
  - Creating and validating models
  - Privacy, inclusiveness, benefiting all citizens





# Tiramisu Bus Tracker App

- People can see incoming bus data
- People can also share info
  - Got on bus
  - #seats available
- Can we create new kinds of tools that can engage people to be active citizens?



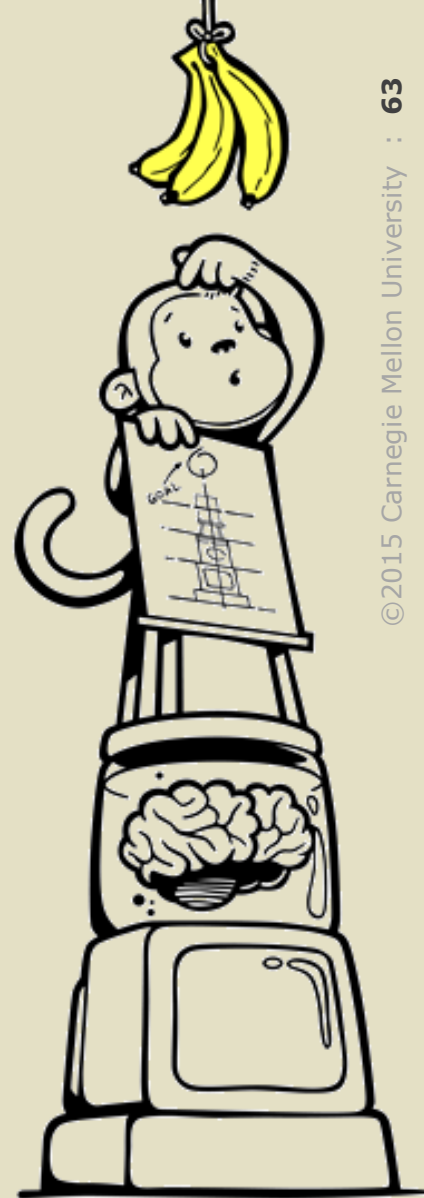
# Thanks!

Would love to hear your feedback and ideas!

More info at [cmuchimps.org](http://cmuchimps.org)  
or email [jasonh@cs.cmu.edu](mailto:jasonh@cs.cmu.edu)

Special thanks to:

- Justin Cranshaw
- Dan Tasse
- Hyun-Ji Kim
- Emily Su
- Jennifer Tchou



Computer Human Interaction:  
Mobility Privacy Security  
<http://cmuchimps.org>





Smart energy systems analyse  
and optimise electricity consumption  
in commercial buildings

Retail stores that adjust to  
people's age and gender

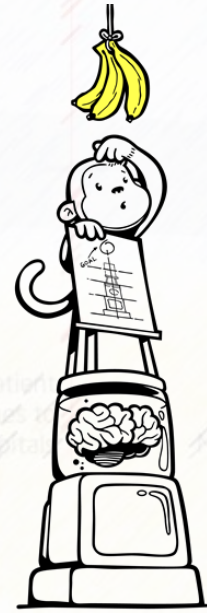
Intelligent traffic systems  
guide drivers away from  
congested roads

# How can we create a connected world we would all want to live in?

**Carnegie  
Mellon  
University**



• **Human-  
Computer  
Interaction  
Institute**



**Computer  
Human  
Interaction:  
Mobility  
Privacy  
Security**

# How Much Can Be Inferred?

MORE  
RISK



Chrome-  
skull accessories

-----  
were in the top 1 percent of  
products signaling a risk  
of default among 85,000 types of  
purchases analyzed.

LESS  
RISK



Premium  
wild birdseed

-----  
was in the bottom 1 percent of products  
signaling a risk of default among 85,000 types  
of purchases analyzed.

# Who Gains From this Data?

- Today, most data only flows one way
  - Mainly to advertisers (and NSA)
  - Also banks, insurance, credit cards





# Who Gains From this Data?

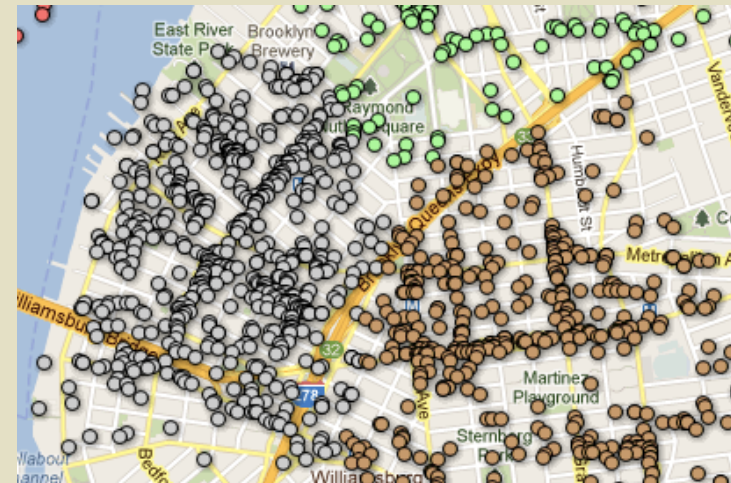
- Can we design systems that share the value across more people?
  - People co-create data **and** gain value
  - Participatory design philosophy
- Can we also make people feel more invested in the cities they live in?



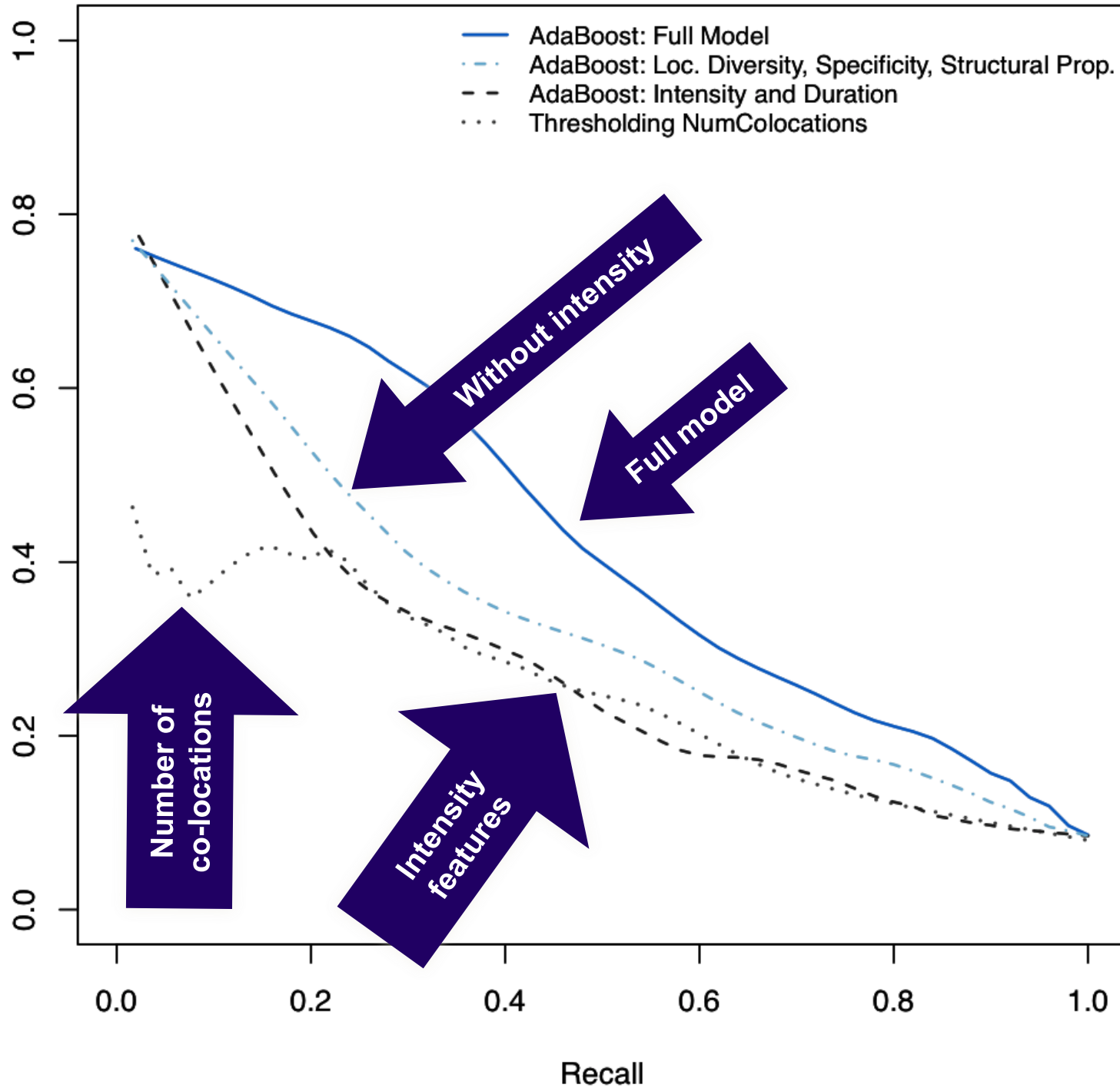


# Summary

- Smartphones and cloud computing offer **big** opportunity to understand human behavior
- Also pose many large challenges, in privacy and ethics
- But I'm optimistic







# Using Location Data to Infer Friendships

- 2.8m location sightings of 489 users of Locaccino friend finder in Pittsburgh
- Place entropy for inferring social quality of a place
  - #unique people seen in a place
  - $0.0002 \times 0.0002$  lat/lon grid,  $\sim 30\text{m} \times 30\text{m}$



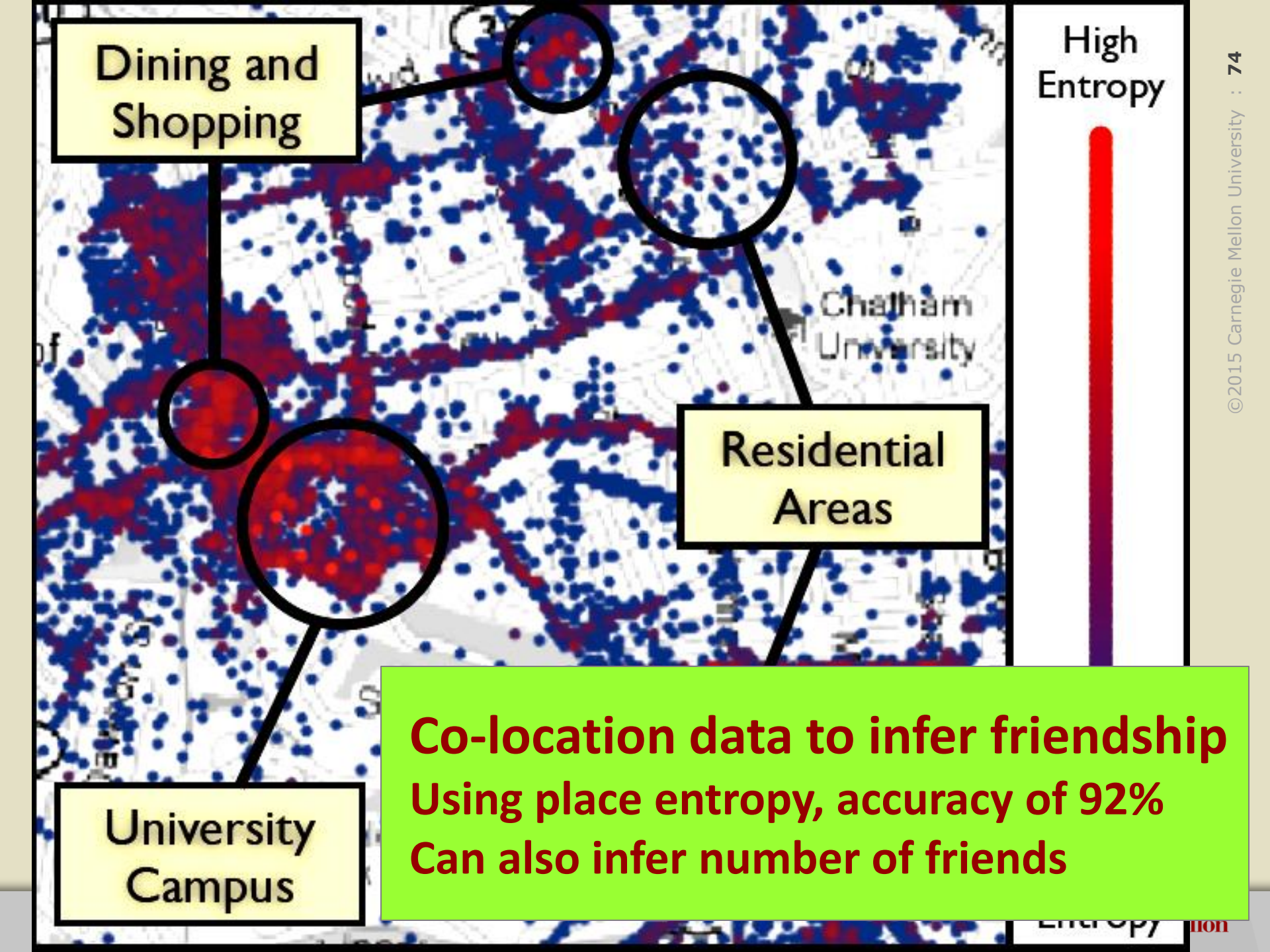
Cranshaw et al, Bridging the Gap Between Physical Location and Online Social Networks, Ubicomp 2010

# Inferring Friendships

- 67 different machine learning features
  - Location diversity (and entropy)
  - Intensity and Duration
  - Specificity (TF-IDF)
  - Graph structure (overlap in friends)
- 92% accuracy in predicting friend/not
  - Location entropy improves performance over shallow features like #co-locations







Dining and Shopping

High Entropy

Residential Areas

University Campus

**Co-location data to infer friendship**  
Using place entropy, accuracy of 92%  
Can also infer number of friends



