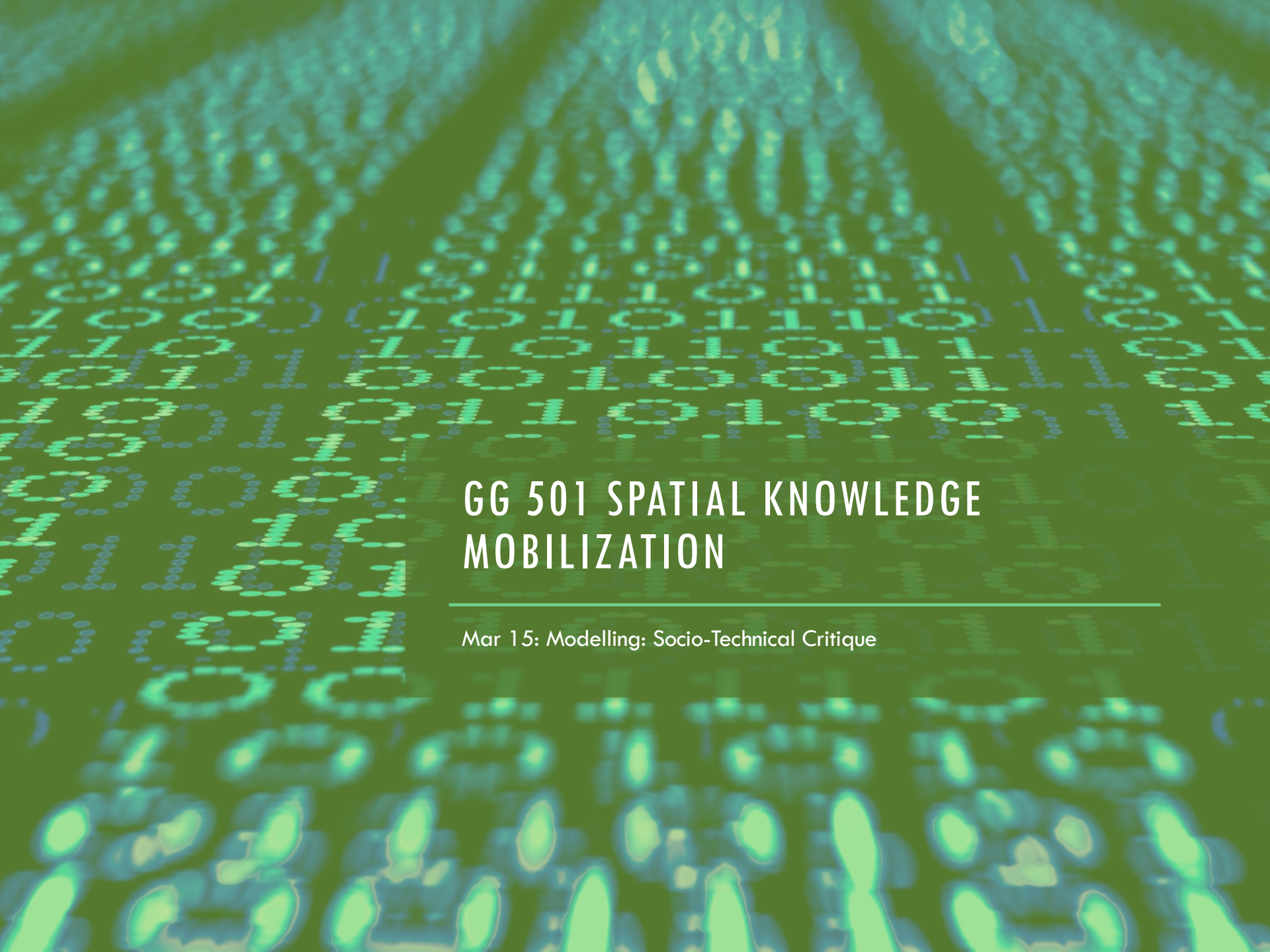


GG 501 SPATIAL KNOWLEDGE MOBILIZATION

Mar 15: Modelling: Socio-Technical Critique

WHAT MAKES A GOOD DATA ANALYST?

Mark Tenenholtz @marktenenholtz · 12h

Every top-tier data scientist is fantastic at:

- Evaluating their models
- Engaging stakeholders
- Building minimally-complex solutions
- Iterating rapidly

Doesn't matter what company/field you're in.

All of them have these qualities.

If you don't, work on them!

3 38 227

QUESTIONS TO ASK WHEN EVALUATING DATA-INTENSIVE RESEARCH

- How modelling/analysis methods relates to research questions
- Stakeholders / end users for the analysis – were they engaged/consulted?
- Model parsimony / KISS
- Data – what types of data were used
  - how was data collected
  - sources of bias
  - sample size?
  - data quality?
    - errors in the data
    - gaps
    - is it current?
    - spatial /geographic coverage complete?
- Model accuracy / validation –
  - how was model validated?
- How were variables selected?
- How were paramaters set?
- Is it reproducible?
  - is the code available, is the data available, is methodology transparent
- Are the conclusions supported by the analysis results?

Big Data and Environment

ACU Summer School
Wilfrid Laurier University

Dr. Colin Robertson
Geography & Env Studies
Wilfrid Laurier University

http://www.thespatiallab.org/

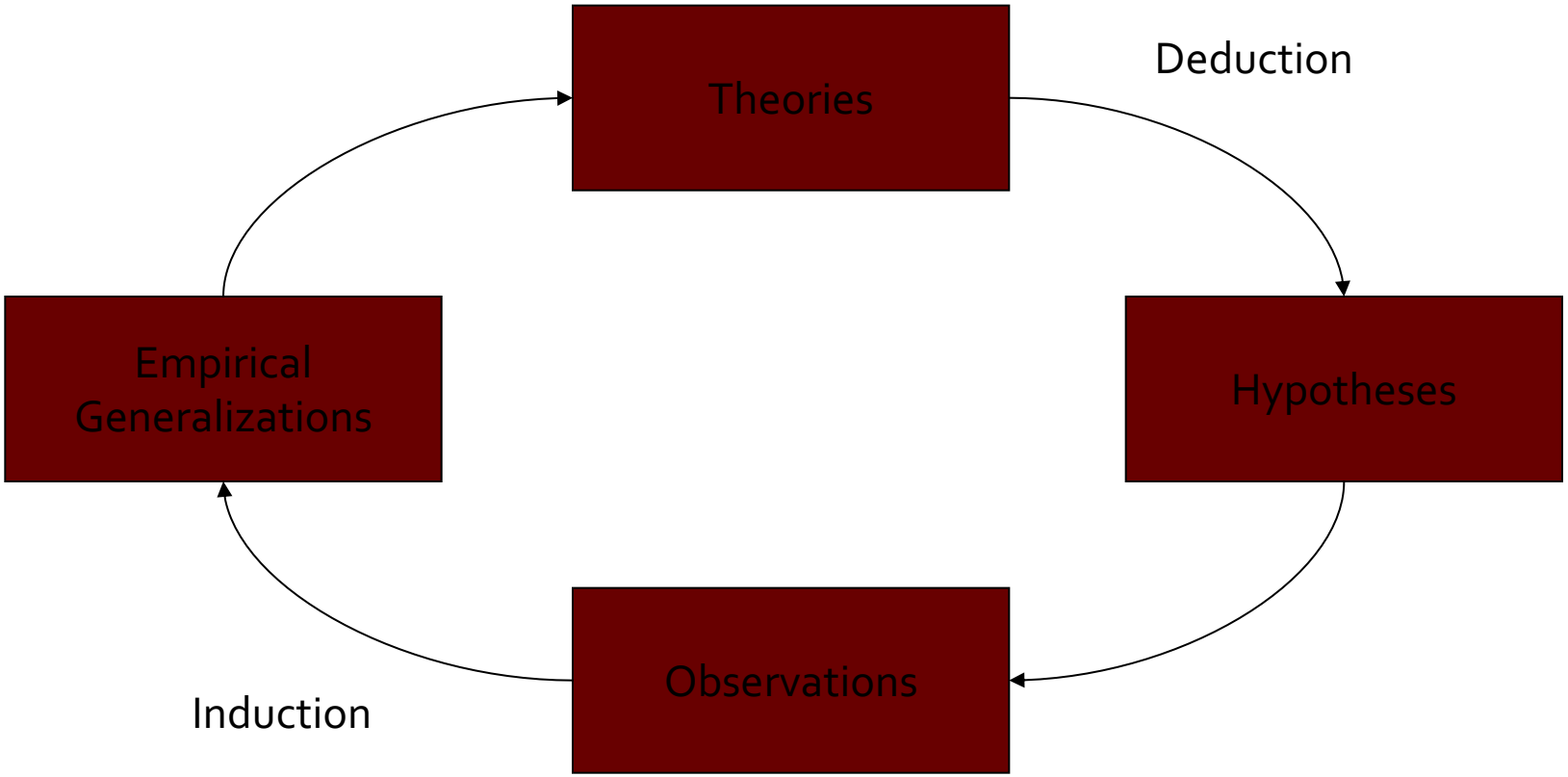
LAURIER
Inspiring Lives.

Intro to Big Data for Environment

What is Big Data?

- Term used to describe new era of data abundance
- Five V's of Big Data:
  1. Volume – Data too large to handle with traditional databases
  2. Veracity – Trustworthiness – huge variety in data quality/formats
  3. Velocity – Speed at which data is generated and moves around
  4. Variety – Different types of data (not just tables, 80% is unstructured)
  5. Value – is it worth the effort?! Need skills and tools to derive value from new sources of data

The Science Process



WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

Illustration: Marian Bantjes

THE PETABYTE AGE:
Sensors everywhere. Infinite storage. Clouds of processors. Our ability to capture, warehouse, and understand massive amounts of data is changing science, medicine, business, and technology. As our

"All models are wrong, but some are useful."

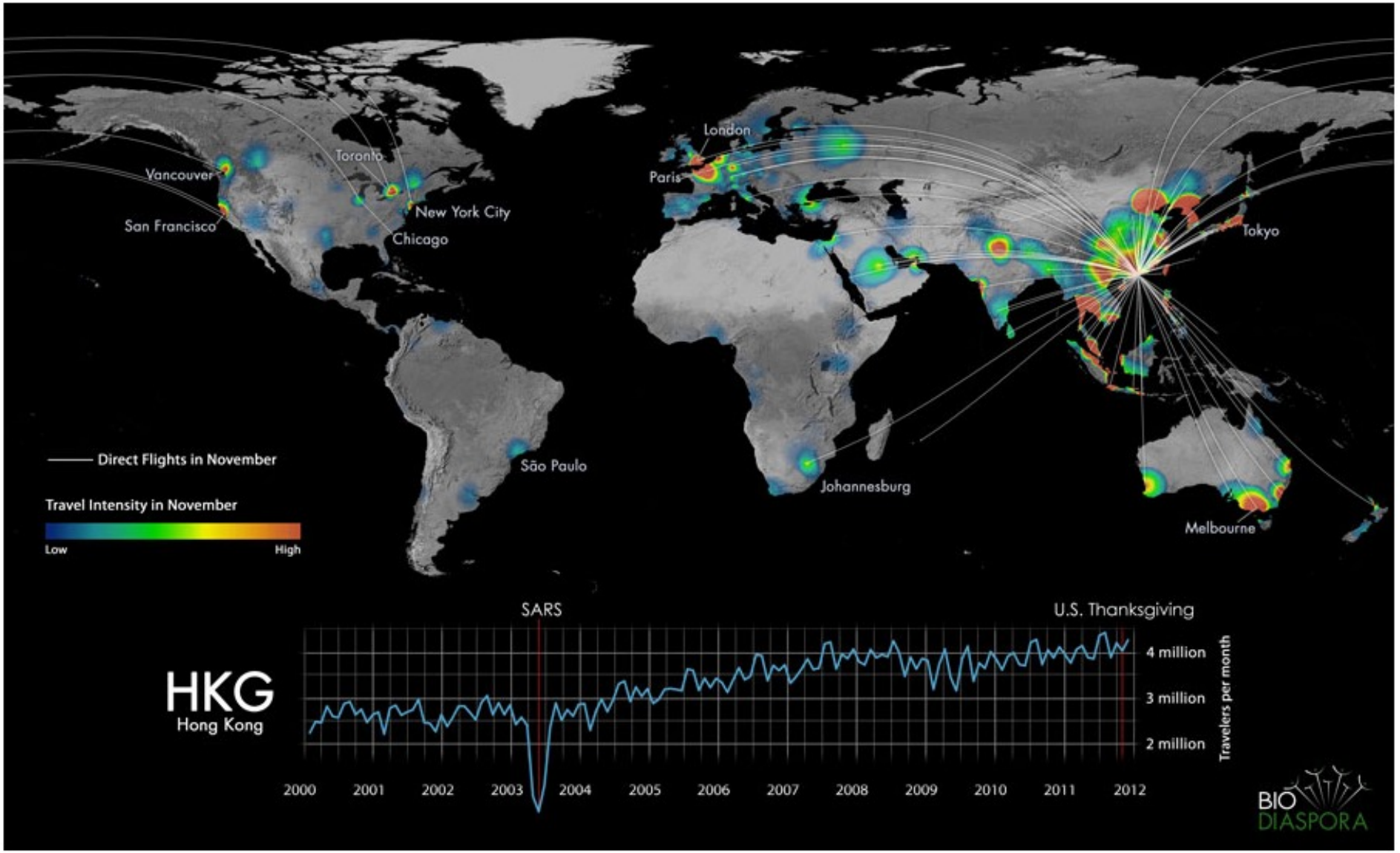
So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human

Where do observations come from?

- Observations are the items that make up DATA of all forms
- Data are defined by a **type**, and measured with some degree of **precision**
- Precision needed for data is dependent on the required knowledge need
- Question to ask: is the data precision adequate for the purpose at hand (also true for data quality)

Conclusions from data

- All of us need to draw conclusions from our experiences
- Increasingly are experience of the world is 'digitized"
  - Implications for society
  - Implications for learning from this data by drawing conclusions from previous observations
- We use data analysis tools and statistical modelling to draw conclusions from data

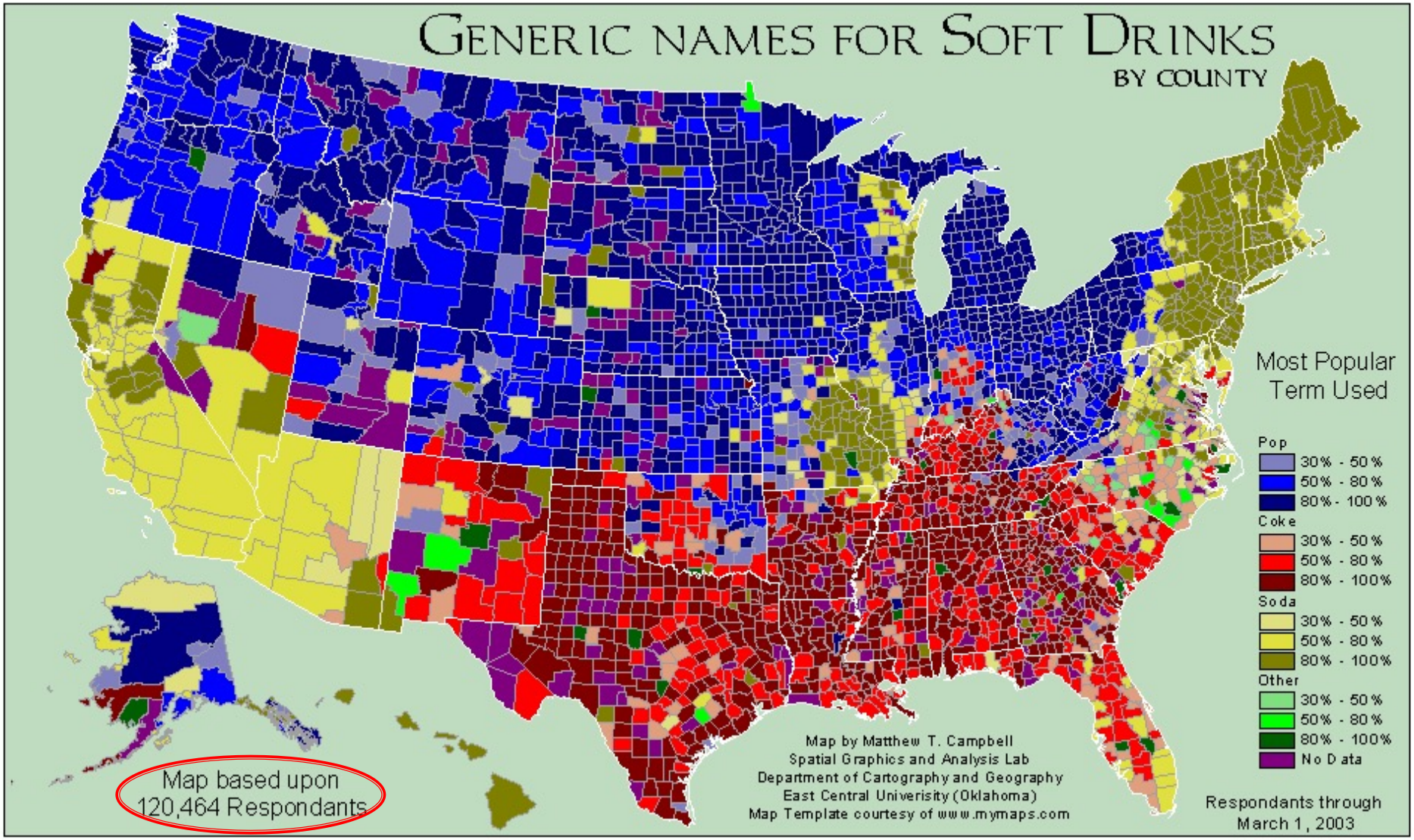


Global air travel as a conduit for the worldwide spread of infectious diseases

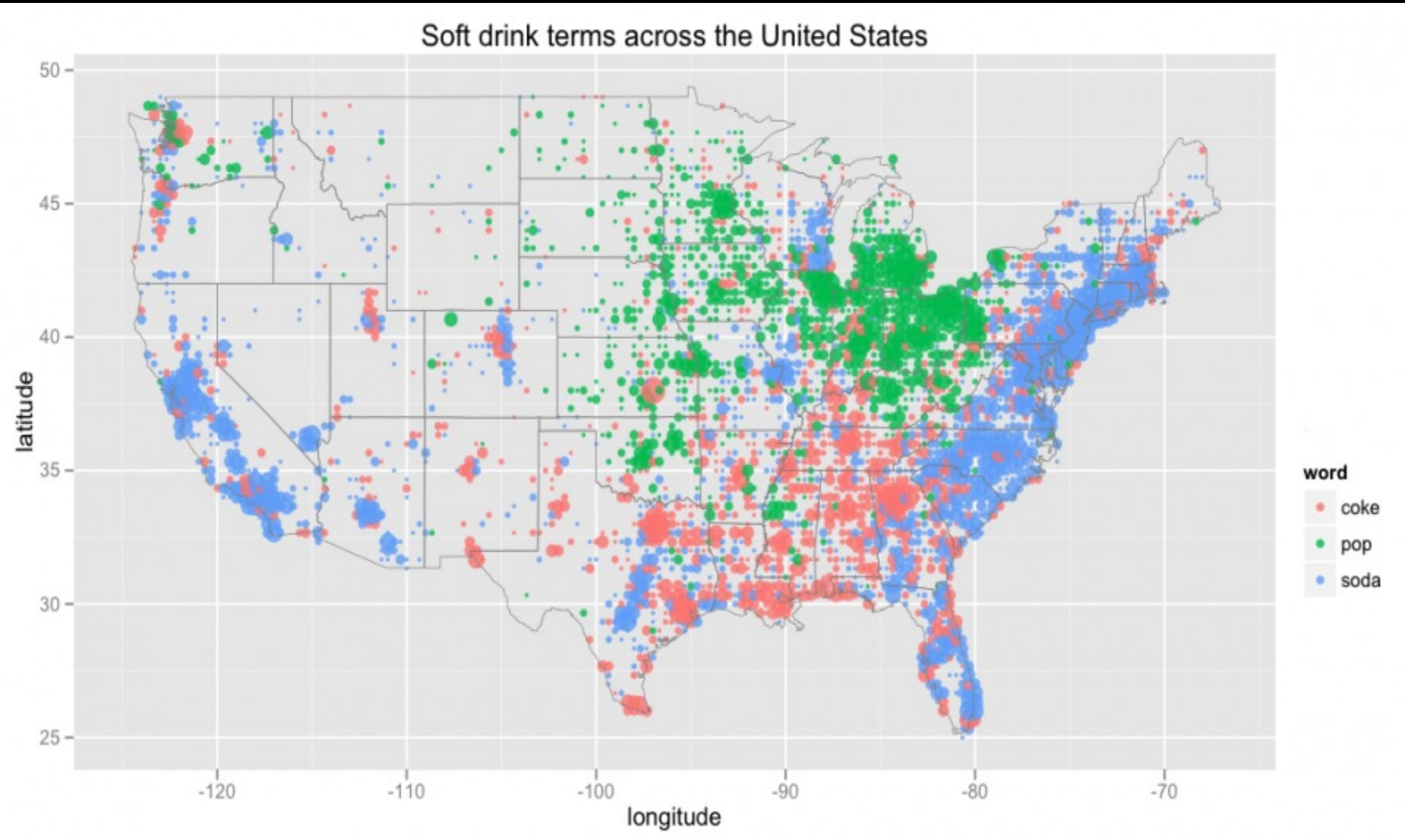
CLOSE ✕

http://www.biodiaspora.com/

People and Pop - before



People and Pop - now: visualizing geotagged tweets



Overview of Today

1. Big data for Environment – Research examples
2. Critical big data
3. Environmental Big Data – Remote Sensing
4. Case Study – Unmanned Aerial Vehicles for Environmental Data Collection

Defining Big Data

- Big data has been hard to define
- Examples
  - transactional data
  - social media data
  - cell phone data
  - satellite imagery?
  - crowdsourcing
  - sensor networks
  - DIY?

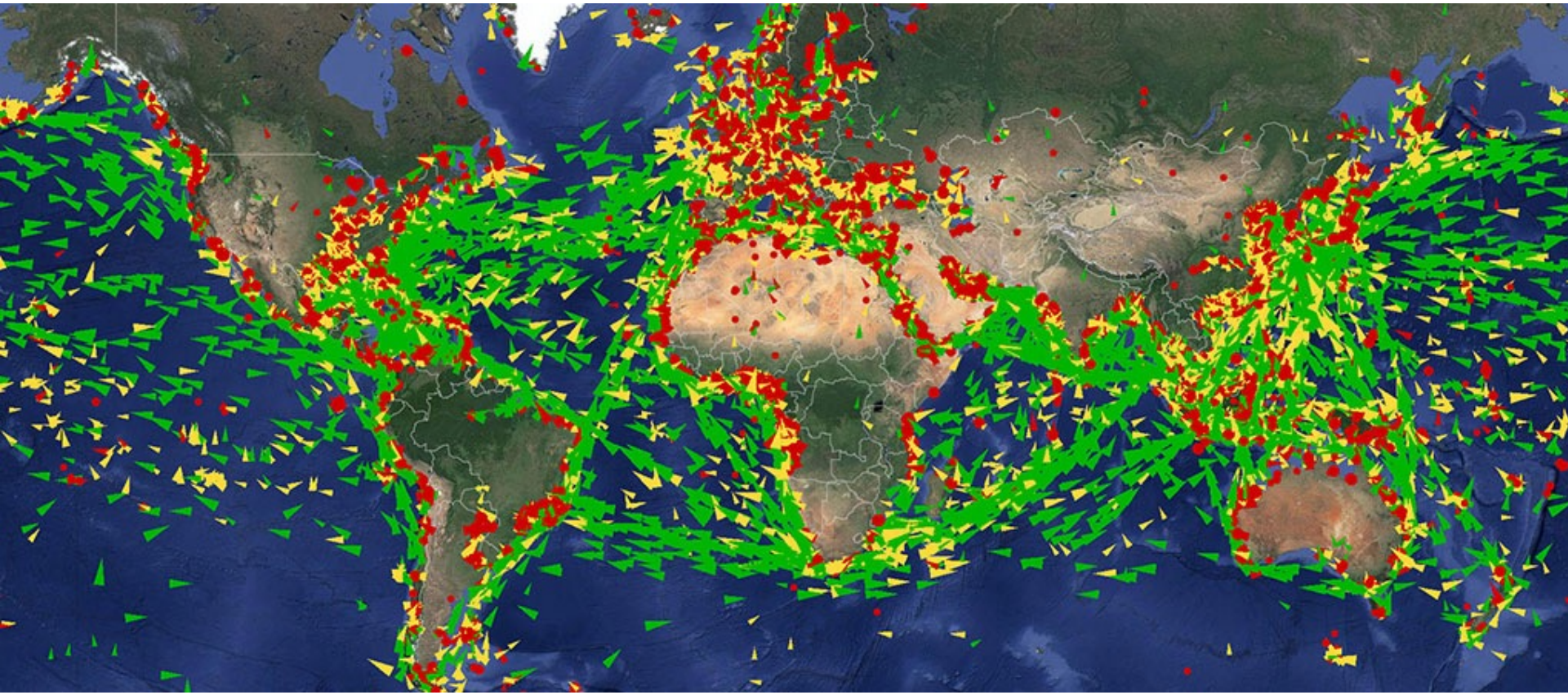
Research Samples using Big Data for Environmental Research

1. Global Shipping Data

Global Shipping Patterns

- Automated Identification System (AIS) sensor on satellite used for locating ships at sea with AIS broadcasters
- AIS messages include location and additional info
  - Cargo
  - Speed / heading
  - Ship details: class, engine specs
  - Trip details: origin, destination, home port

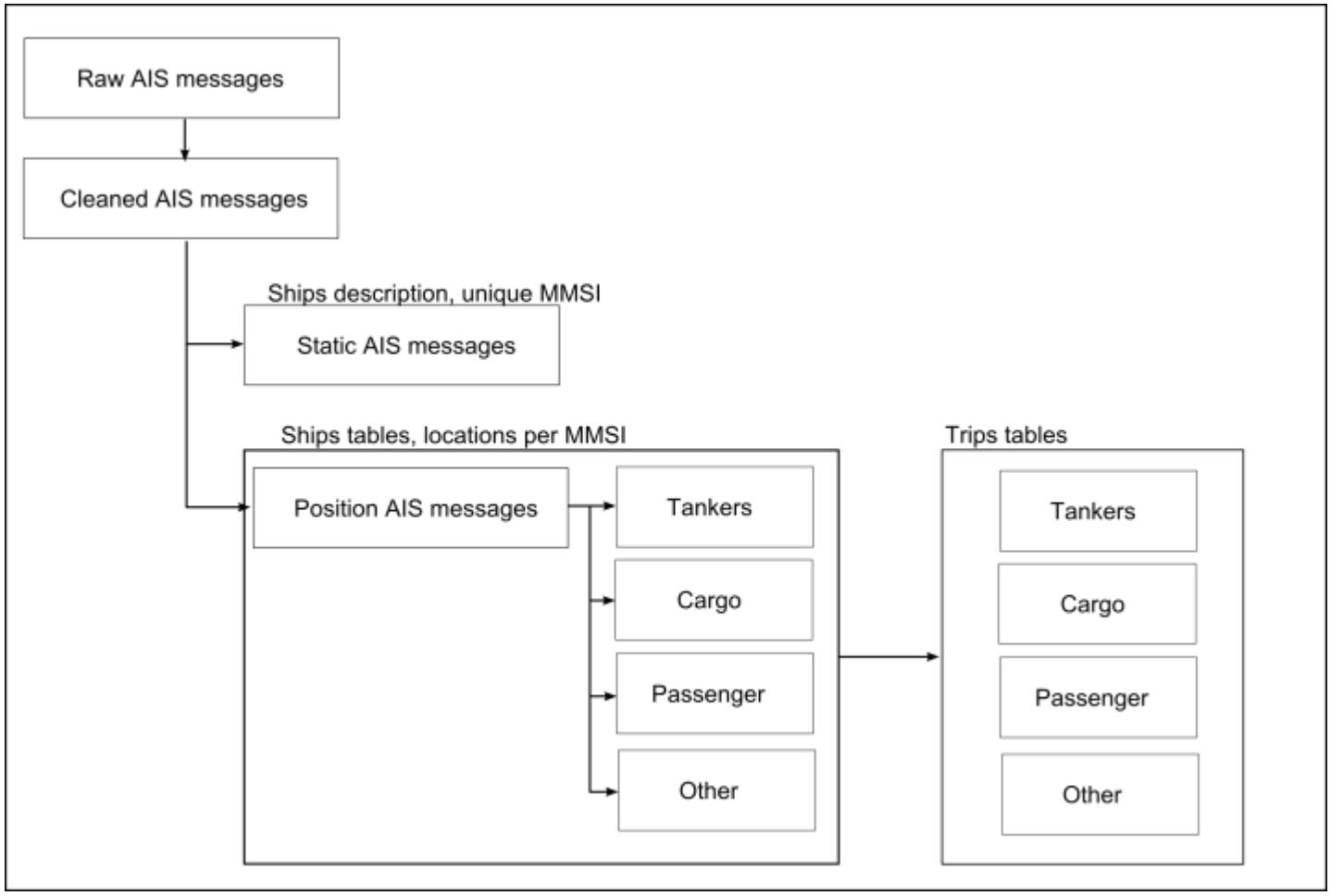
AIS Data is BIG Data



AIS Data

- Each message has raw location data
  - Latitude and longitude
- And a time stamp
- Need to construct trips out of raw points

AIS Data


Raw AIS messages
Cleaned AIS messages
Ships description, unique MMSI
Static AIS messages
Ships tables, locations per MMSI
Position AIS messages
Tankers
Cargo
Passenger
Other
Trips tables
Tankers
Cargo
Passenger
Other

What can we do with this data?

- Track country to country shipping traffic
  - trade activity etc.
- Number and origin of shipping companies for any given port
- Cargo load / type. etc
  - distribution of trade for different cargo types
- Shipping routes for oil tankers
  - oil spill risk?
- illegal activities
  - illegal fishing
  - human trafficking
  - refugee / illegal migration

What questions should we ask about the data?

- how accurate is the data?
  - how often is it update
  - how many ships actually have the AIS transponder
    - is it easy to turn off
  - how accurate is the spatial/location data
- completeness of attributes
- bias:
- spatial gaps / temporal gps

Extracting Ship Trajectories

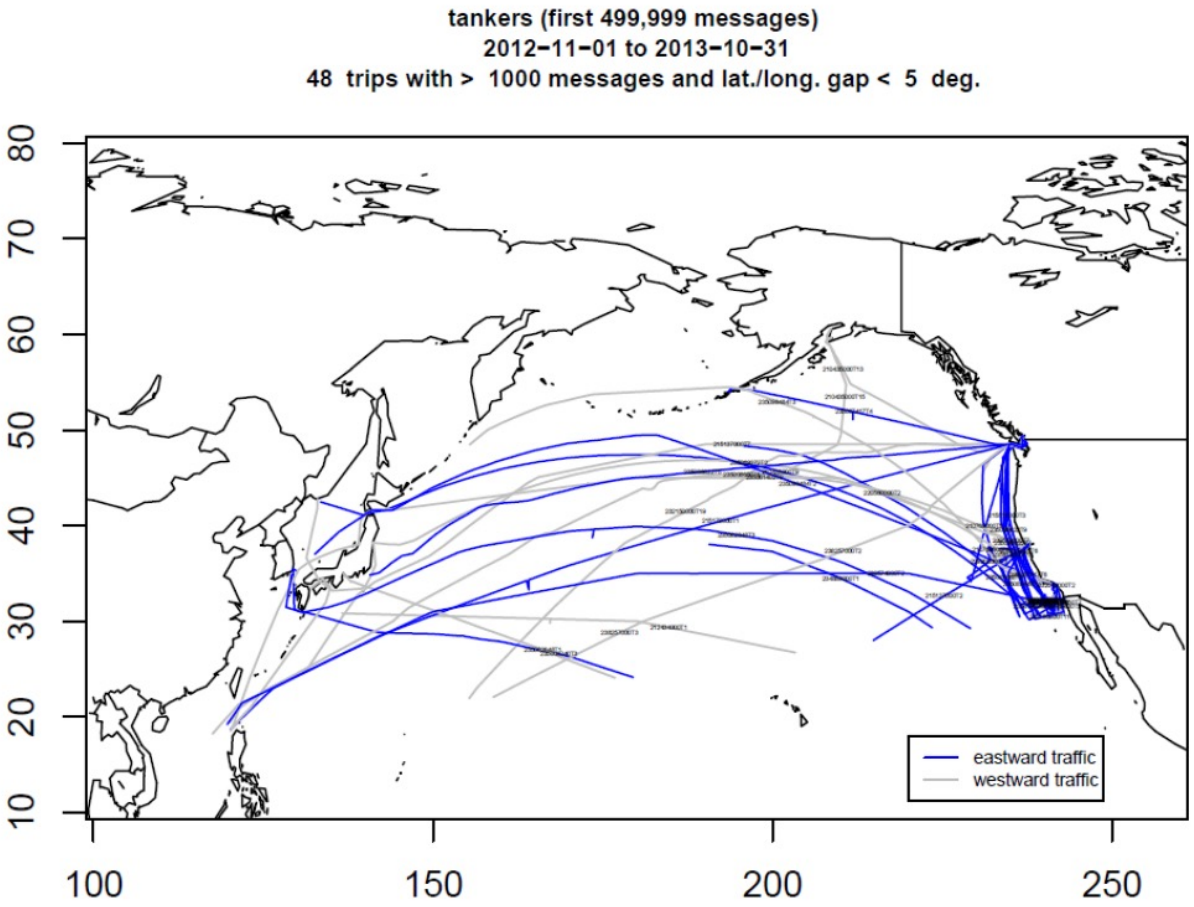


Figure 2 Trips extracted from raw AIS messages based on navigation status, trips with greater than 1000 messages and latitude and/or longitude gaps less than 5 degrees.

Extracting Ship Trajectories

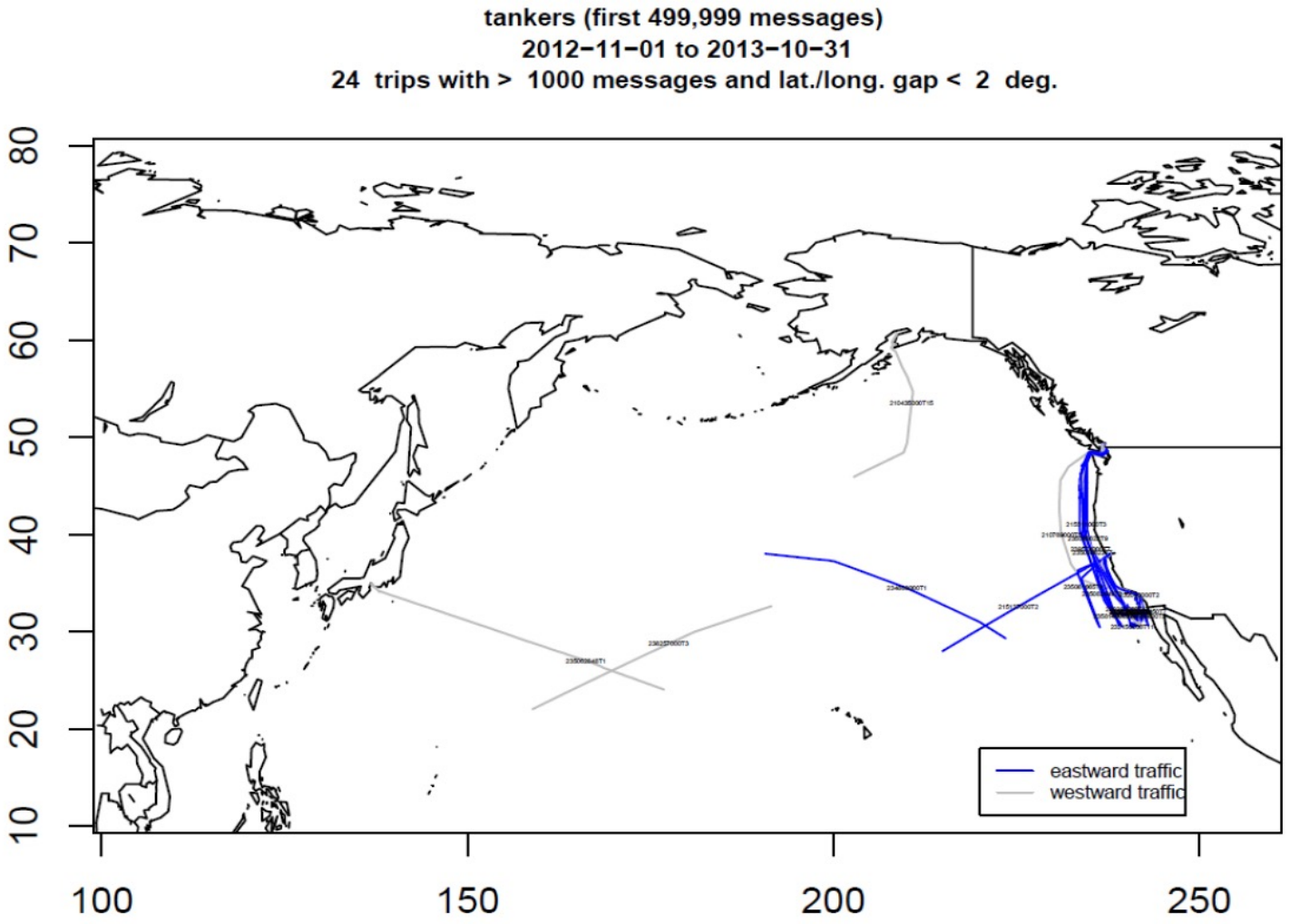
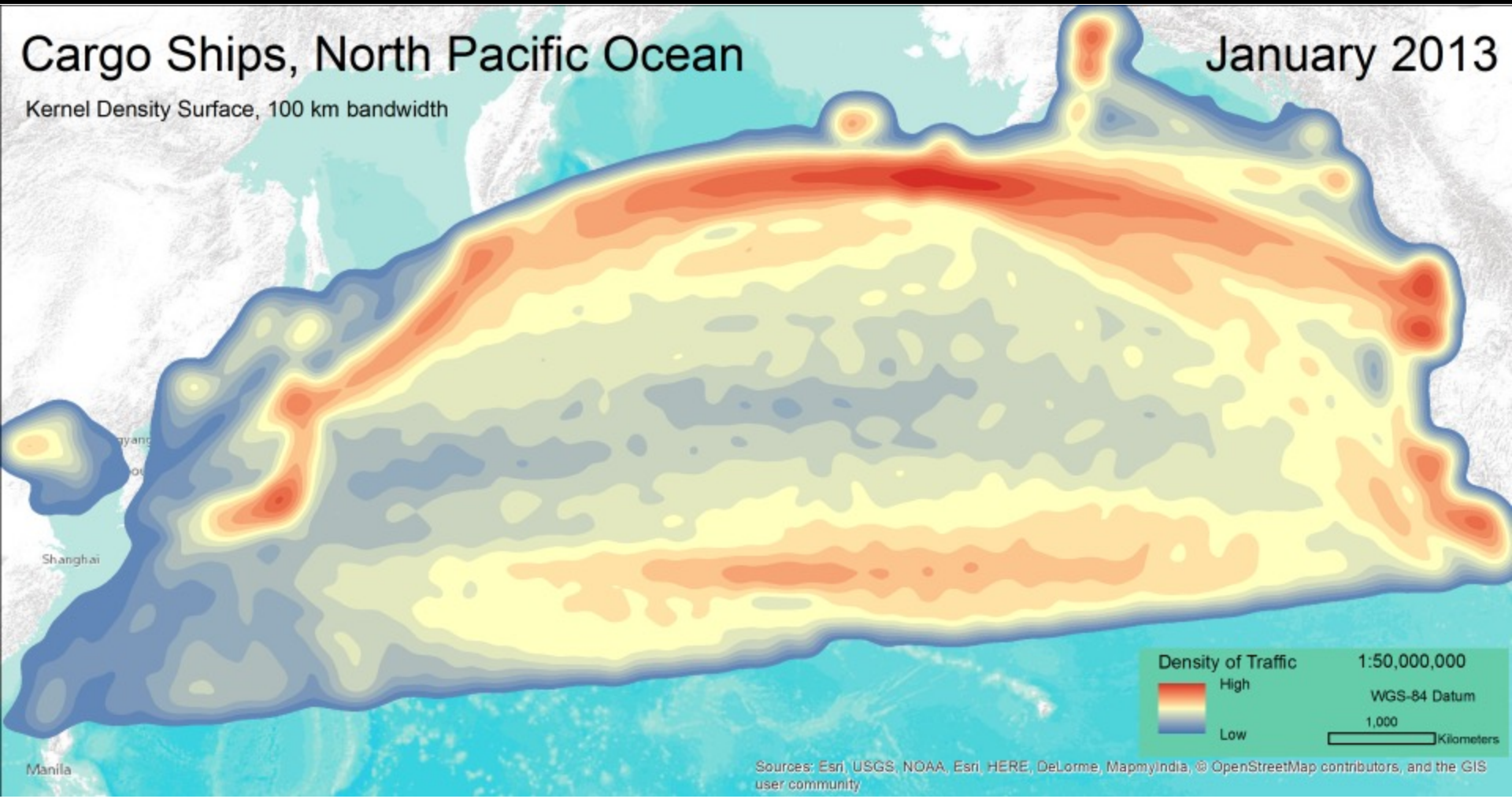
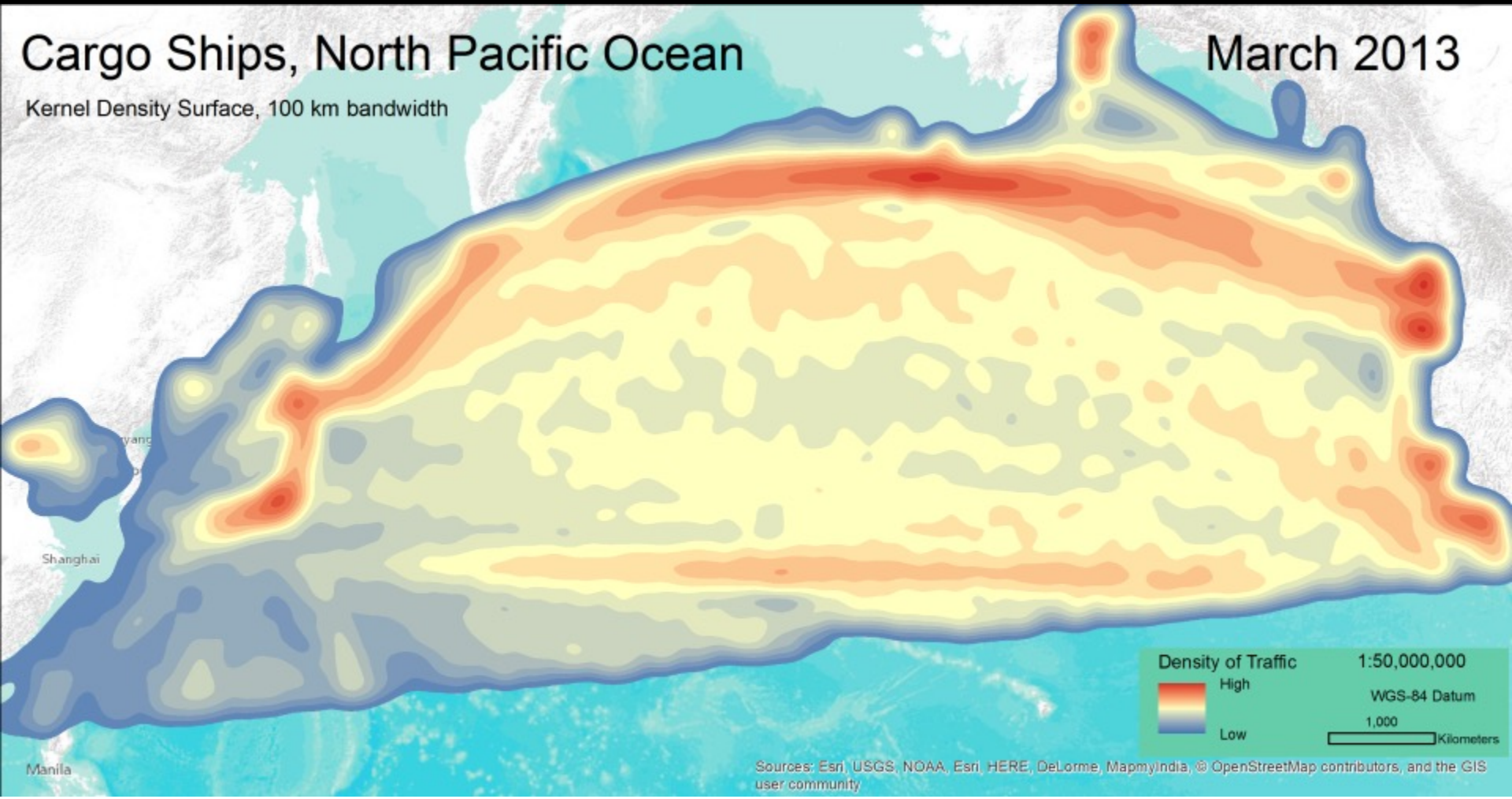


Figure 3 Trips extracted from raw AIS messages based on navigation status, trips with greater than 1000 messages and latitude and/or longitude gaps less than 2 degrees.

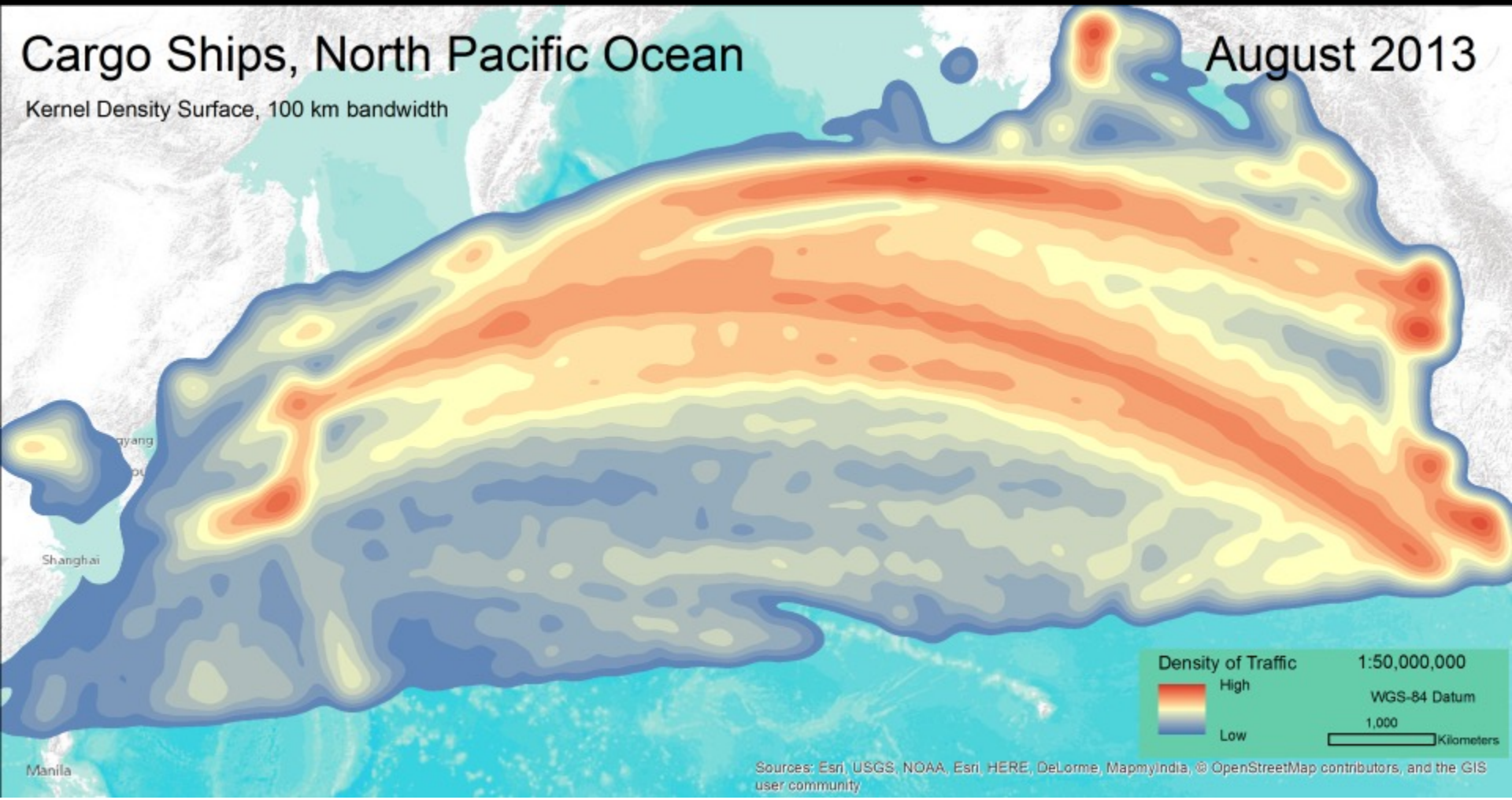
Shipping Patterns



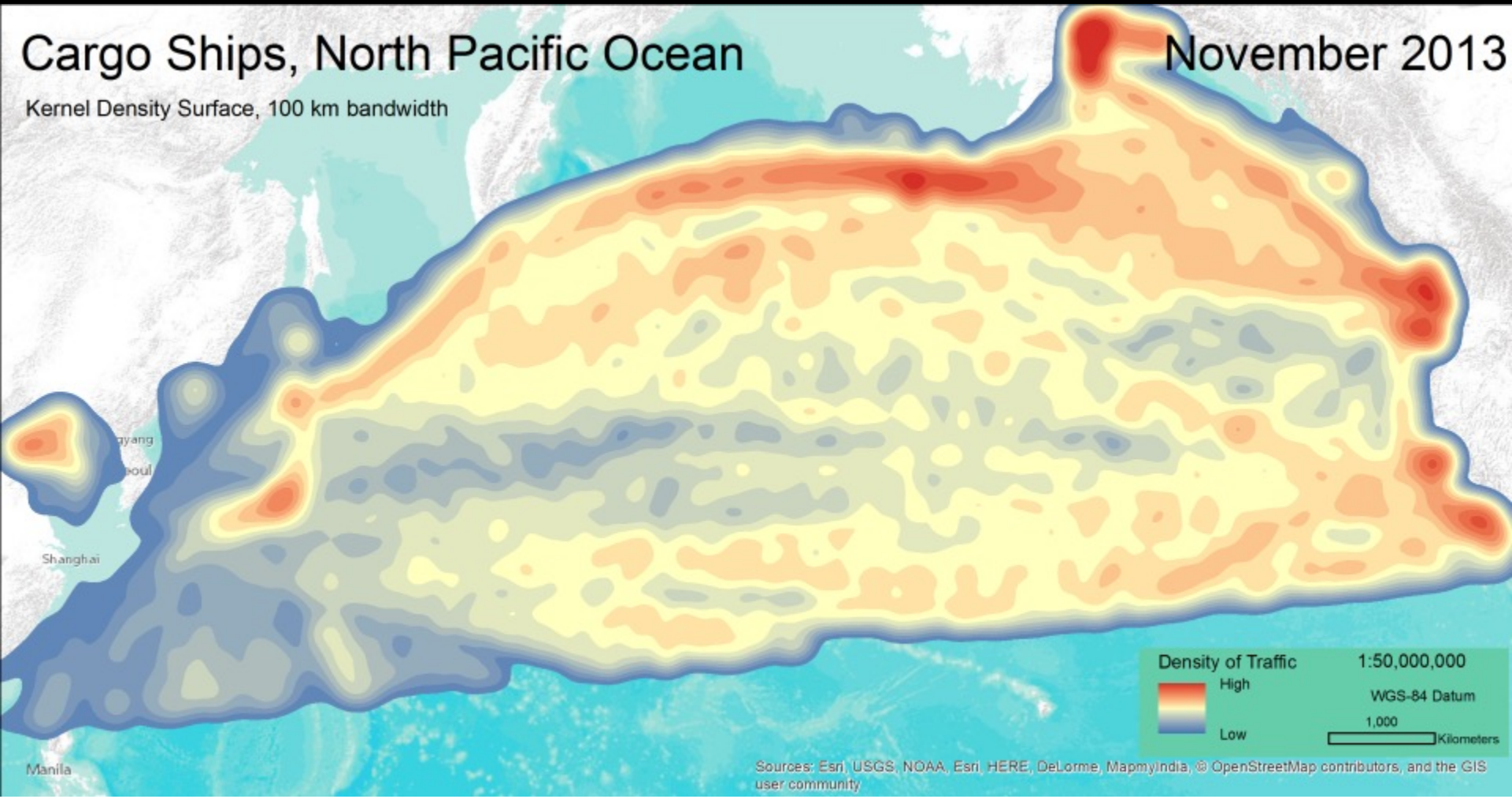
Shipping Patterns



Shipping Patterns



Shipping Patterns



2. GeoSocial Data Mining

The Digital Skin

- Cities are 'information factories' – creating layers of data contouring spatial and temporal rhythms of life
  - digital traces – social media, cell phones, transactional data, smart city sensors
- Rabari and Storper describe the aggregate of these layers as the digital skin – one characterized by
  - uneveness
  - change over time
- Research centres on repurposing derma in the digital skin for examining questions of space and place

i. Geo + Social

- Geographic data
  - data with references to physical location
- Social data
  - data created through social exchange / interactions

Figure 2.5: VGI framework

Scientific knowledge VGI Type 1	Local knowledge VGI Type 2	Personal knowledge VGI Type 3
- volunteered - Objective - Structured - 2 way or n way - Digitizing, GPS, twitter - Only points or points, lines, polygons	- facilitated-VGI - Subjective - Unstructured - 1 way, 2 way or n way - Selection, Drawing, geocoding - Points, lines, polygons	- Kept private - Subjective - Unstructured - n way - Location through networks, geocoding - Only points

Deparday 2010

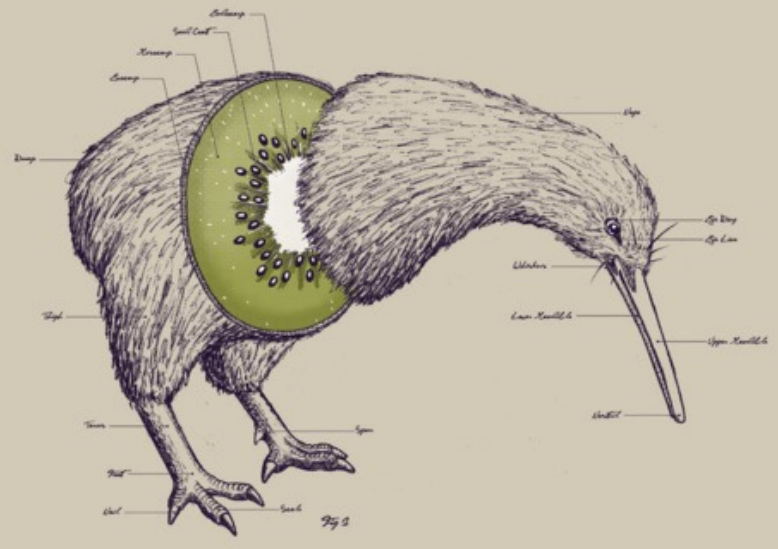
ii. Spatial Pattern Analysis

- Spatial processes create patterns
- Characterizing spatial patterns can reveal underlying processes
  - relation to social / environmental context
- GIS → space as index for data integration
- Open data, open APIs, increasing spatial sources available for integration

iii. Natural Language Processing

- The best nuggets are always hardest to mine!
- Big Data in social sciences is often unstructured
  - decoding meaning
  - recognizing and managing space and time references
  - disambiguation!!
- The things humans do well, computers do poorly

Kiwi (disambiguation)

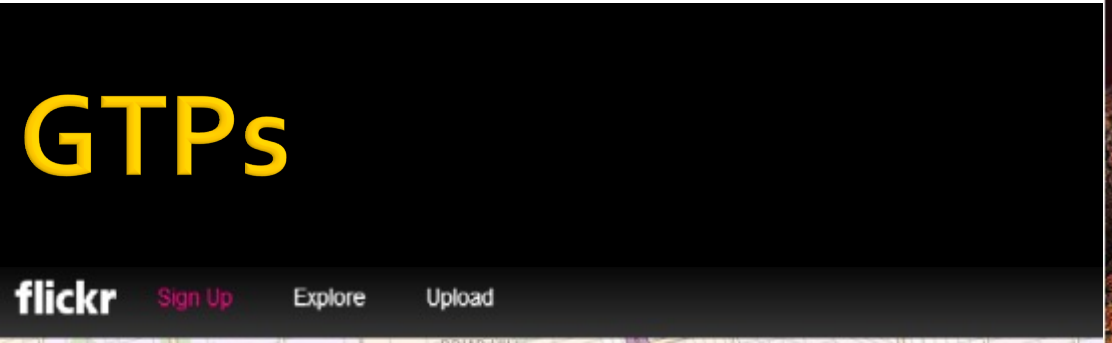


http://www.surf-matic.com/

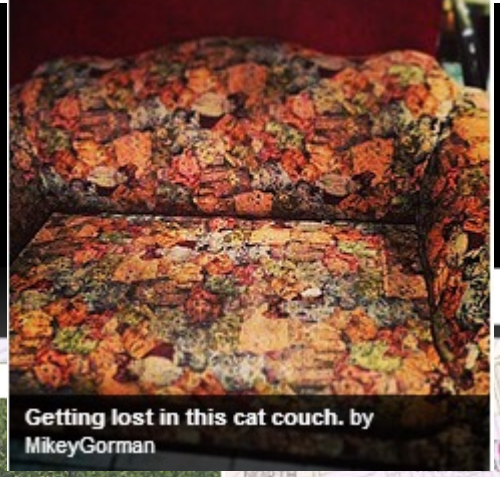
1. Urban Place Mining

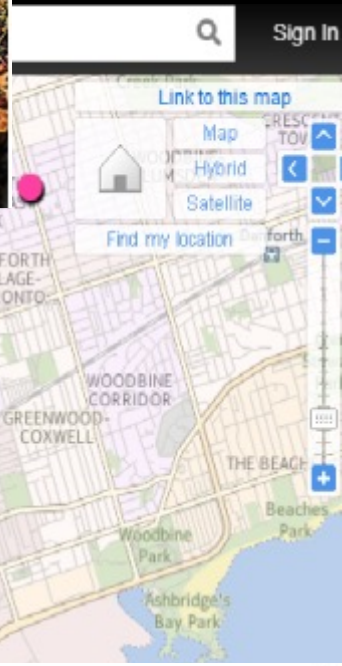
- Geographers are deeply interested in understanding places
- Place – can be defined as "space with meaning"
- Can geo+social data be used to better understand place-making activities
  - And promote the building and nurturing of vibrant urban places

GTPs



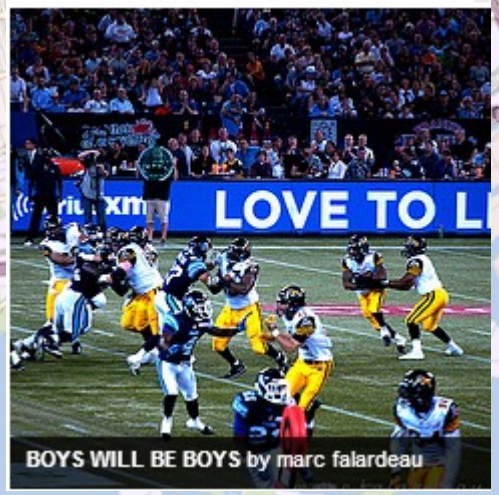
Tagged "cats" in Toronto


Getting lost in this cat couch. by MikeyGorman

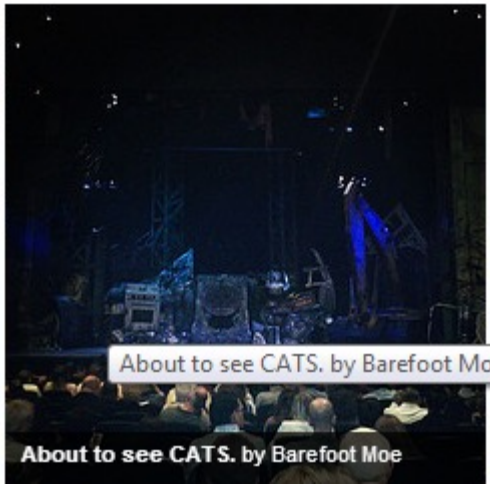



Cierto Azul The story of a blind orphan boy adopted by a jazz sextet consisting only of cats ficción costarricense by josue salazar


day two. my gal likes routines. wait, i must, too. #cat #ritual by sweetie pie press



BOYS WILL BE BOYS by marc falardeau



About to see CATS. by Barefoot Moe

What can we do with this data?

What questions should we ask about the data?

Research Objectives

- Develop methods to explore how urban space and place are characterized in GTPs
- Explore how semantic similarity changes with scale based on GTP tagging
  - Is an area described the same way when GTPs are aggregated?
- Identify areas of agreement and disagreement in how they are characterized in GTP tags.

Study Area

- Vancouver
  - Heterogeneous physical and social landscapes
    - Land use transformation around urban core → the postindustrial city
    - Rapid growth
    - Geographic constraints
    - High density "livable region" urban development plan
    - Cultural & social diversity
      - Recent influx of migrants
  - Tourism – key economic driver

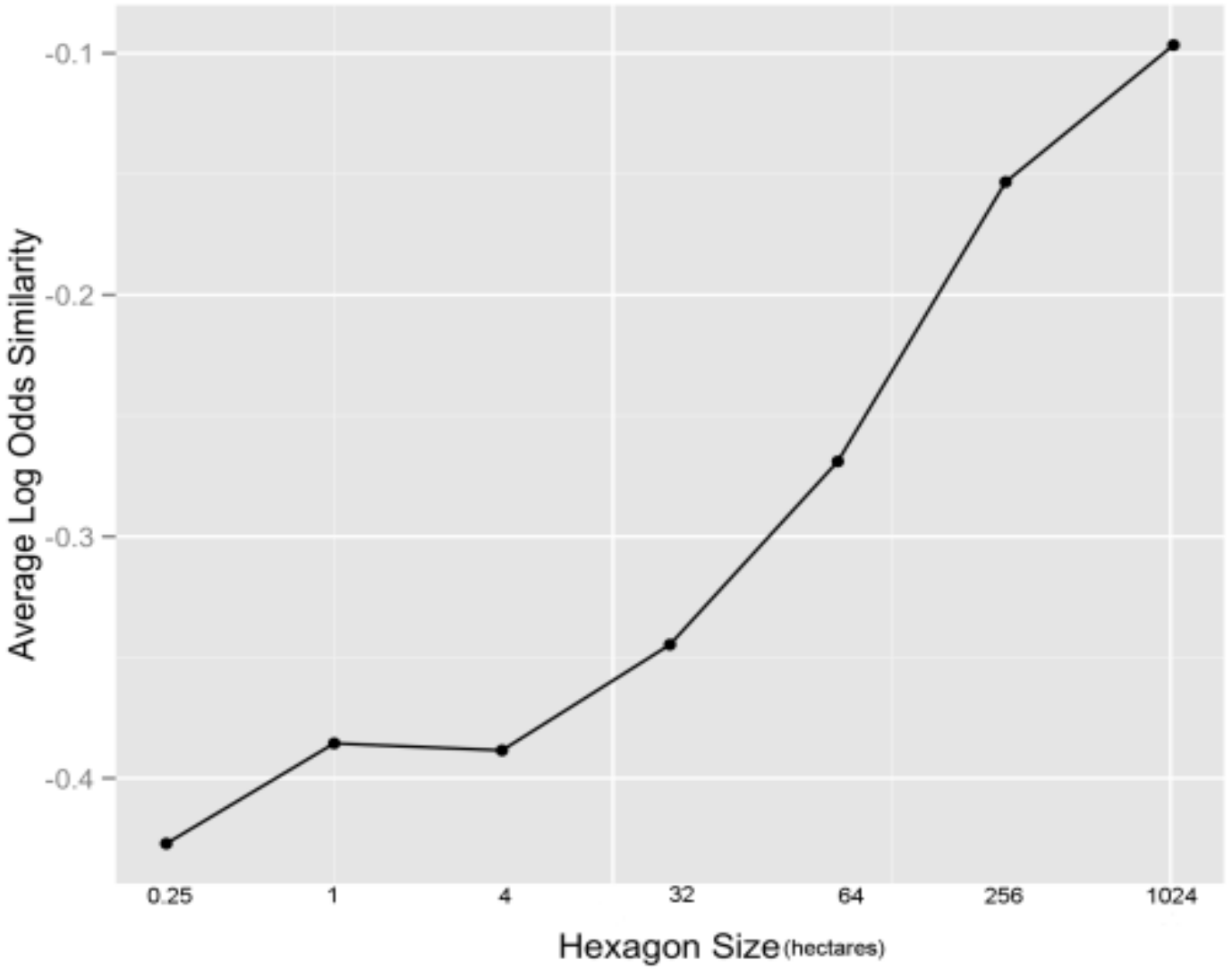
Generating Tag Sets

- Aggregation of Flickr points on 7 hexagon scales
- For each scale-hex combination
  - Generate vector of unique tags and tag counts
  - Constrain vector to 10 most frequently occurring tags
  - Decompose each vector into attributes:
    - Hex1="Vancouver", Hex1Count=500
    - Hex2="BC Place", Hex2Count = 250
    - …
    - Hex10="Butterfly", Hex2Count = 2

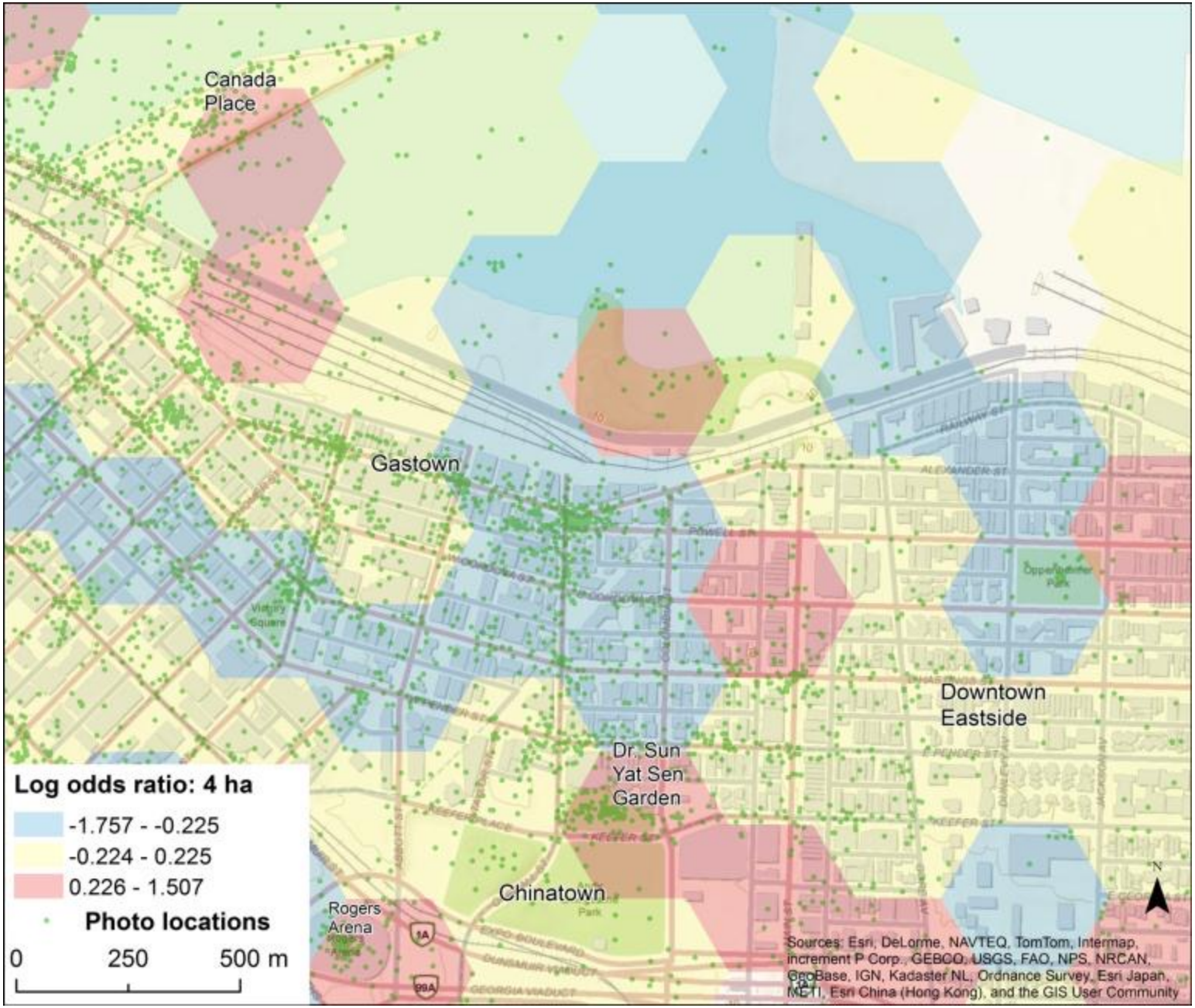
Analysis Methods

- Spatial-semantic similarity
  - Needed to measure how each tag set for a given hexagon compared to its neighbourhood
  - Data comprised of a vector comprised of top10 tag counts for each hexagon
- Proportion of each tags in Hex *i* compared to proportion in neighbouring hexes, across all tags.
- Do spatial neighbours tag the same as you?

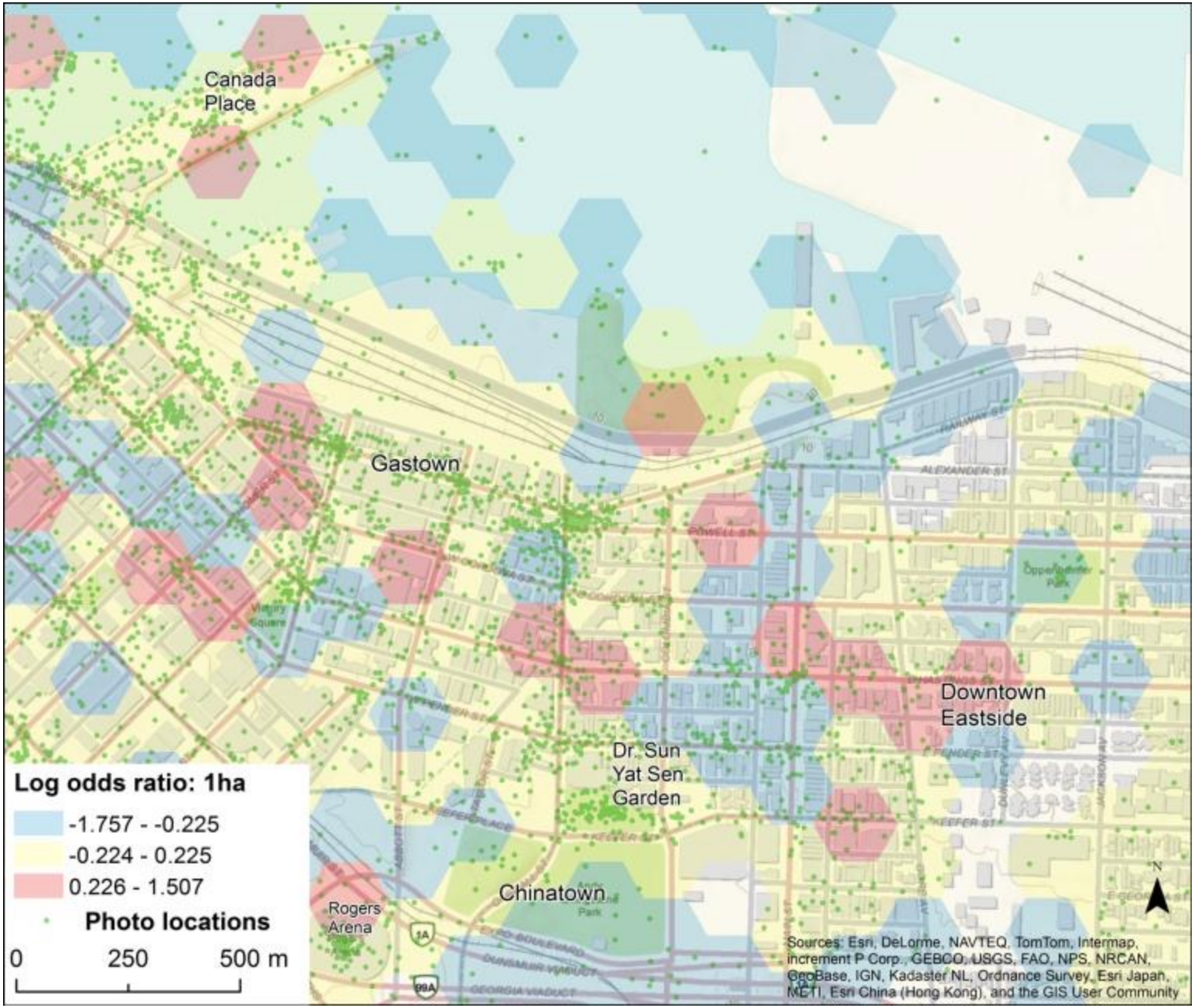
Similarity and Scale



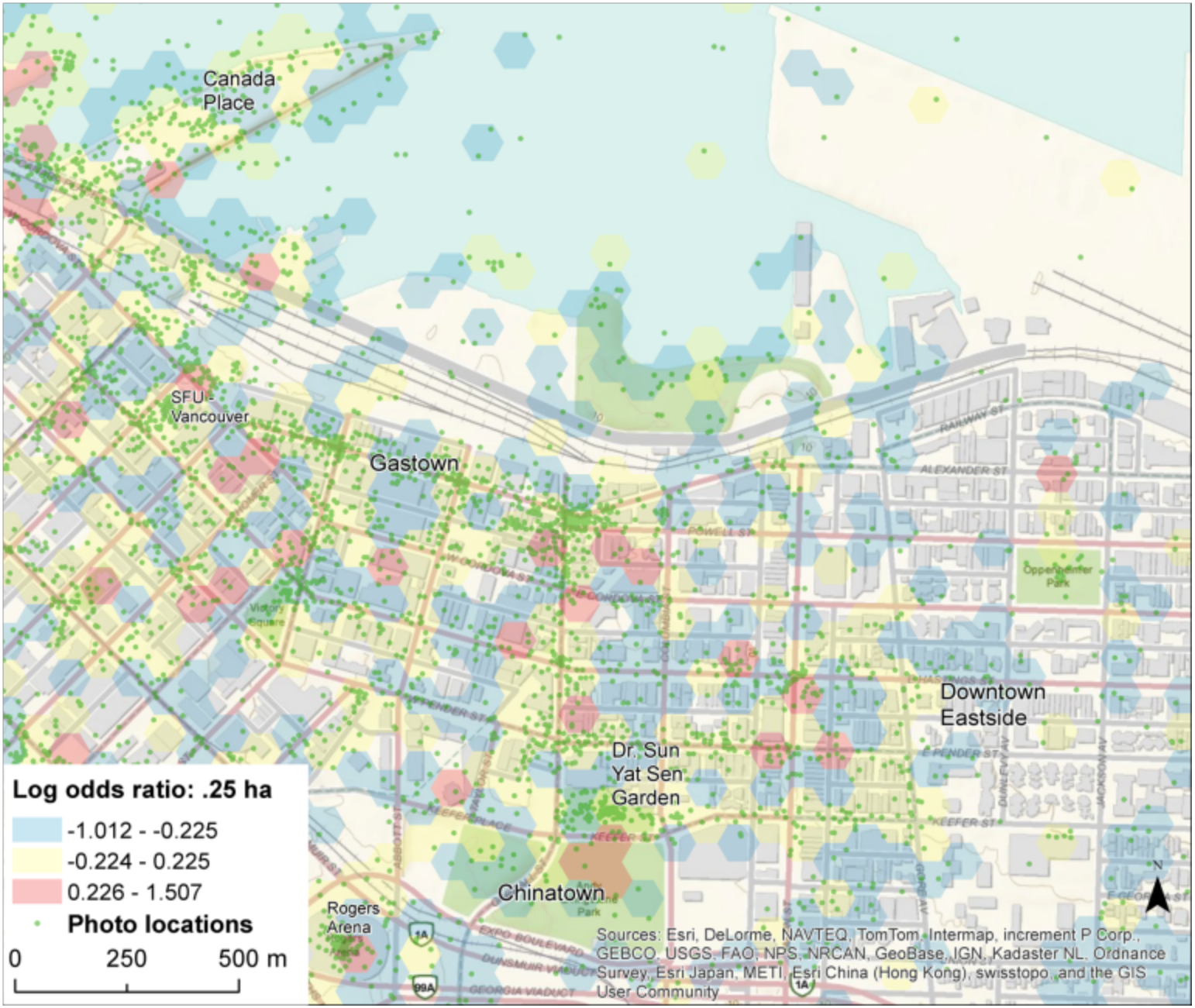
Spatial-semantic Similarity



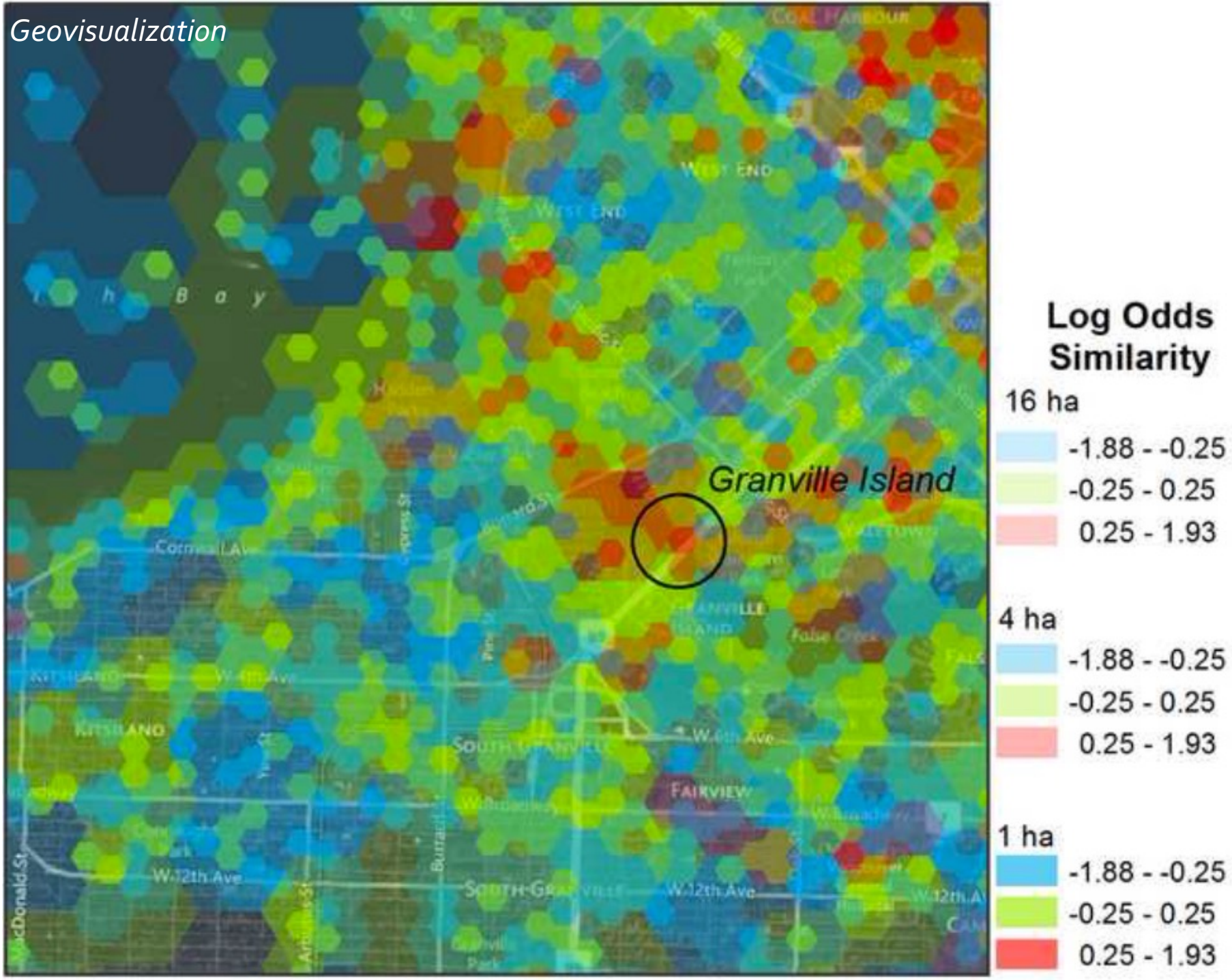
Spatial-semantic Similarity



Spatial-semantic Similarity



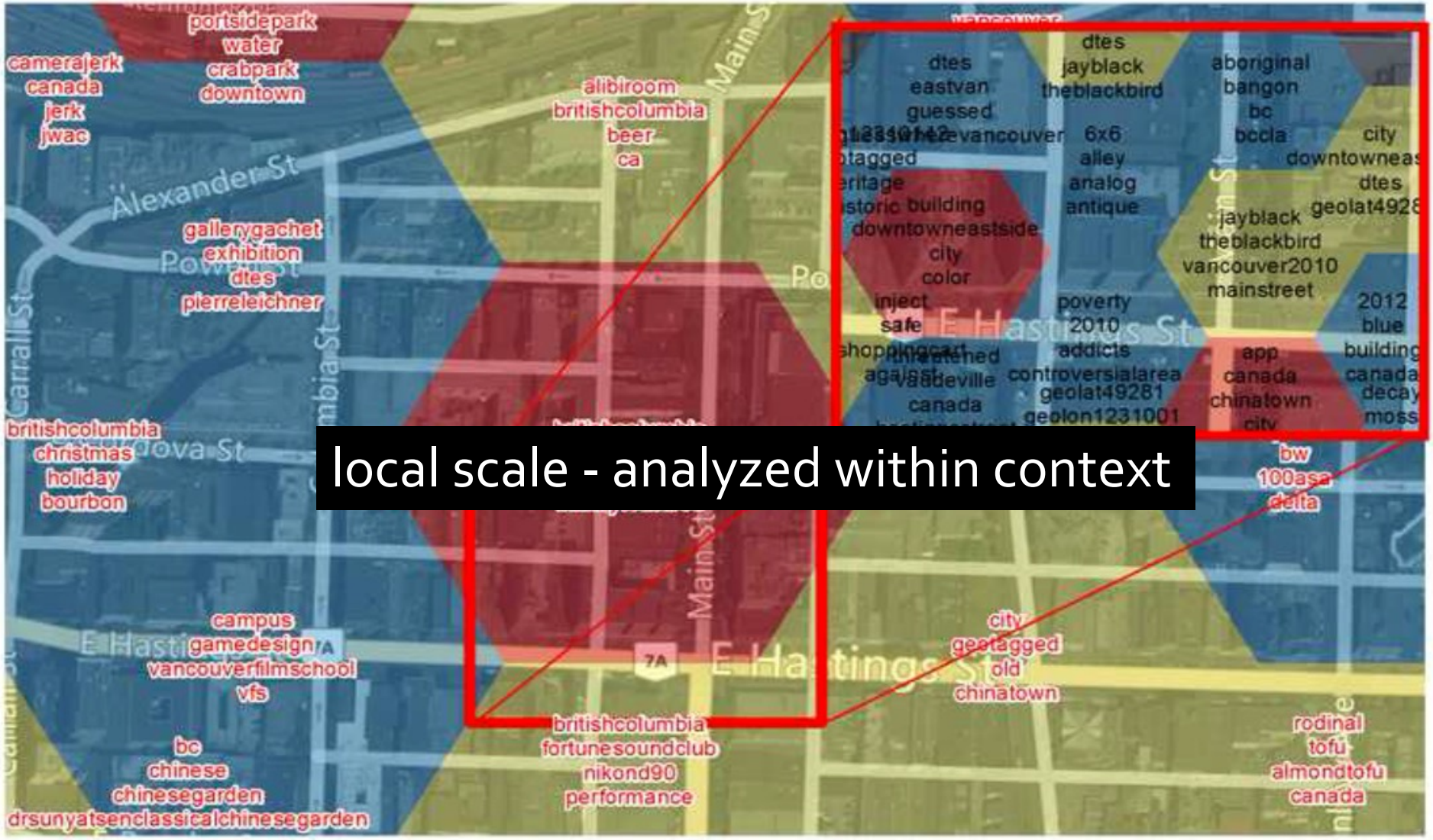
Geovisualization



Tag Frequencies – are boring…

Scale	Most frequent Tag	2nd Most frequent tag	3rd Most frequent Tag
0.25 ha	Vancouver (1699, 16%)	BC (868, 8%)	Canada (844, 8%)
1 ha	Vancouver (1296, 21%)	BC (556, 9%)	Canada (544, 9%)
4 ha	Vancouver (777, 24%)	BC (314, 10%)	Canada (300, 10%)
16 ha	Vancouver (445, 33%)	BC (154, 11%)	Canada (155, 11%)
64 ha	Vancouver (202, 45%)	Canada (70, 16%)	Canada (70, 16%)
256 ha	Vancouver (75, 54%)	Canada (28, 20%)	BC (31, 22%)
1024 ha	Vancouver (22, 54%)	Canada (11, 27%)	BC (12, 29%)

Tag-Space – local matters



2. Geospatial Dimensions of Twitter

- What can geo+social data tell us about human experience through space and time?
- How can we examine relations between expressions in social media?
  - is it worth it?

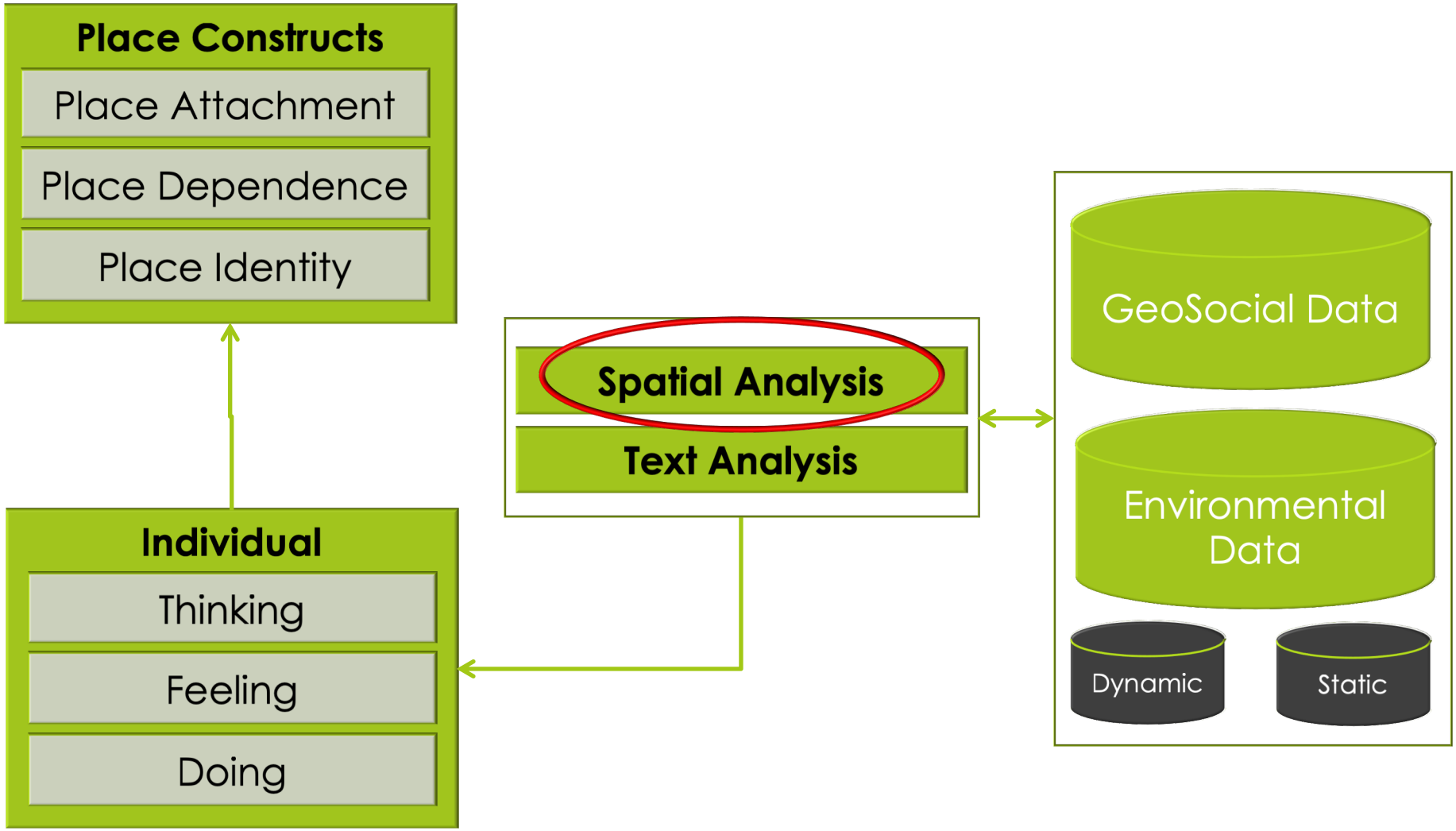
The Health, Happy Digital City

- New science of healthy cities recognizes how our environment shapes our health
- Contextual and compositional factors are both important determinants of health
- Place and space interact to form our experience of the city
  - need to measure these constructs over large areas, time periods, across diverse populations
    - How?

Traditional Measures of Place/Space

- Environmental Psychology
  - instrument: surveys
  - individual focus
  - Quantitative
- Human Geography
  - instrument: observation, surveys, interviews
  - individual and community foci
  - Qualitative
- GeoComputational??
  - instrument: geosocial data
  - individual and community foci
  - Quantitative and Qualitative?

Place Sensing Framework



Research Questions

1. How much spatial variation is there in geotagged Tweets?
2. Can we detect patterns of space-use through at the individual level through analysis of geolocated Twitter messages?
3. How do spatial patterns and space-use derived from geotagged Tweets change over time?

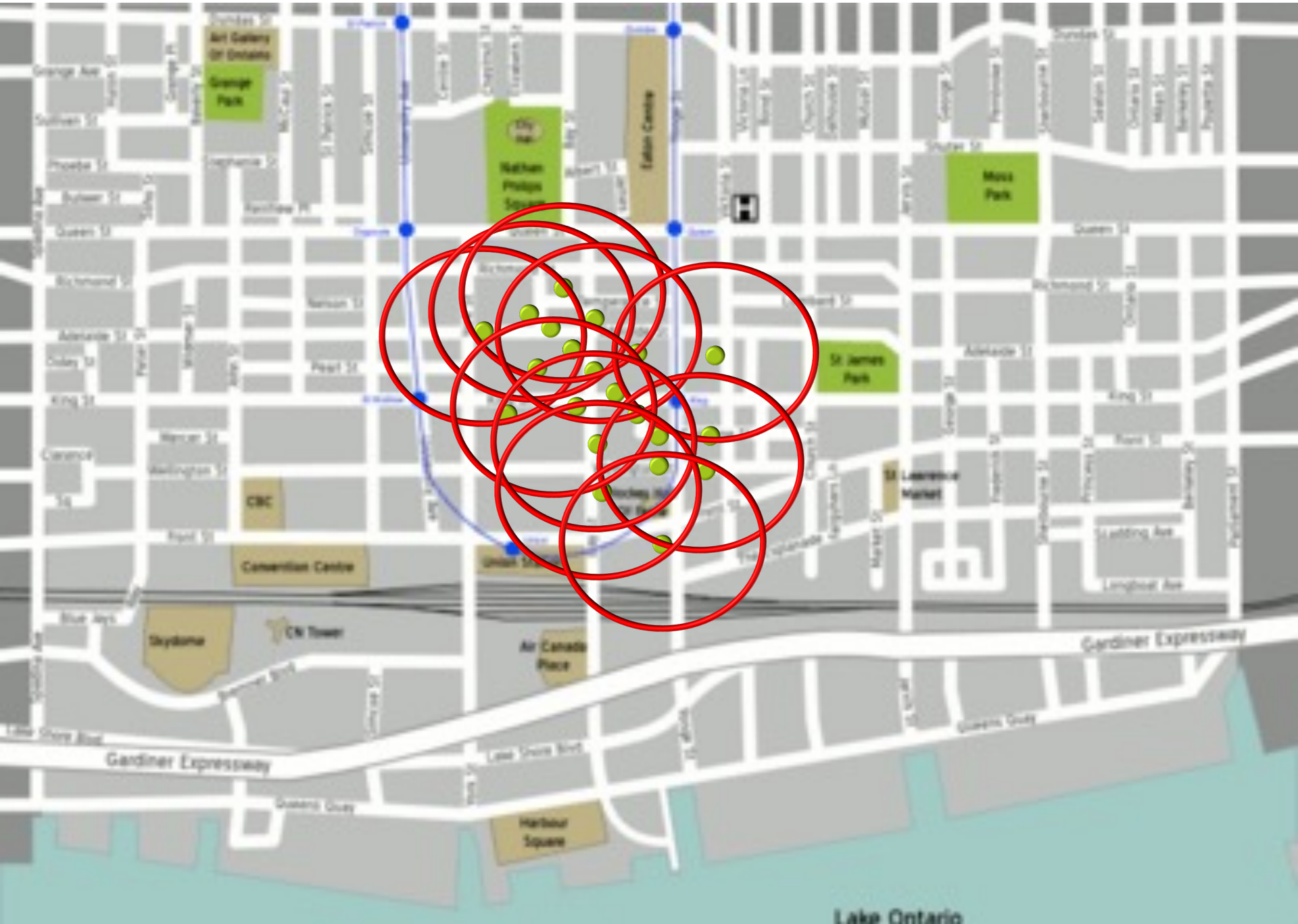
Personal Activity Centres

- Seek to map general areas where people engage in geosocial data production
- Aims are to:
  - reduce data dimension
  - focus on core geographic areas associated with Tweeting
  - associate meaning / function to locations where Tweeting takes place
- Personal Activity Centres: geographic locations where Tweeting occurs most frequently for individuals

Defining Personal Activity Centres

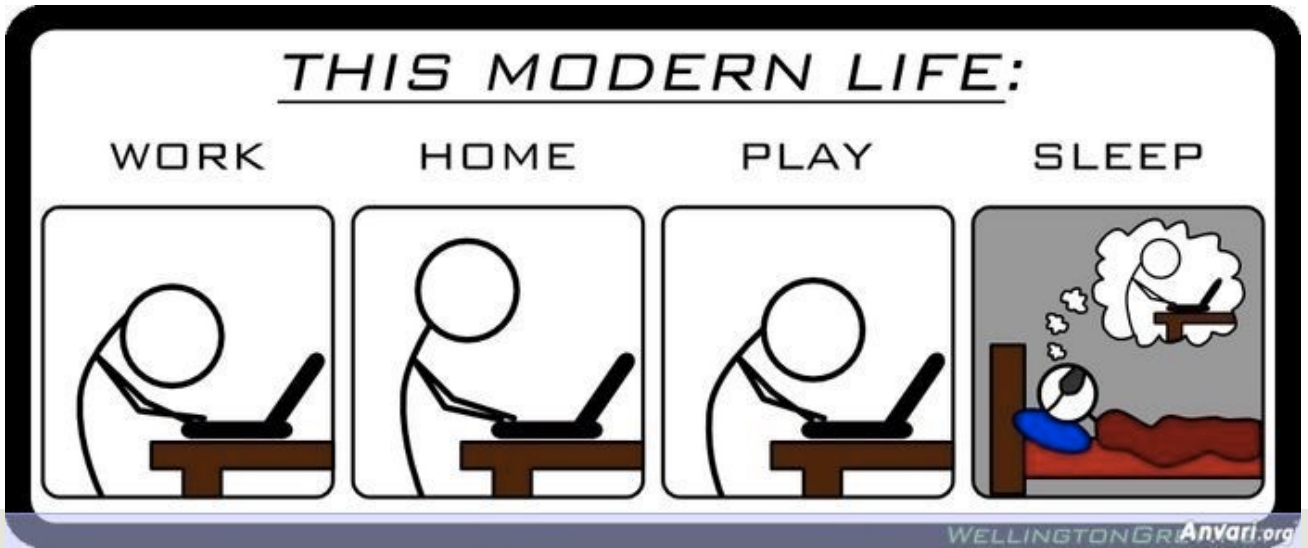
- For each unique user in the data set do the following:
  1. Buffer each tweet by X m
  2. Define a set of activity centres by dissolving boundaries between overlapping buffers
  3. Compute density of Tweets / Area for each PAC
  4. Order ACs by density

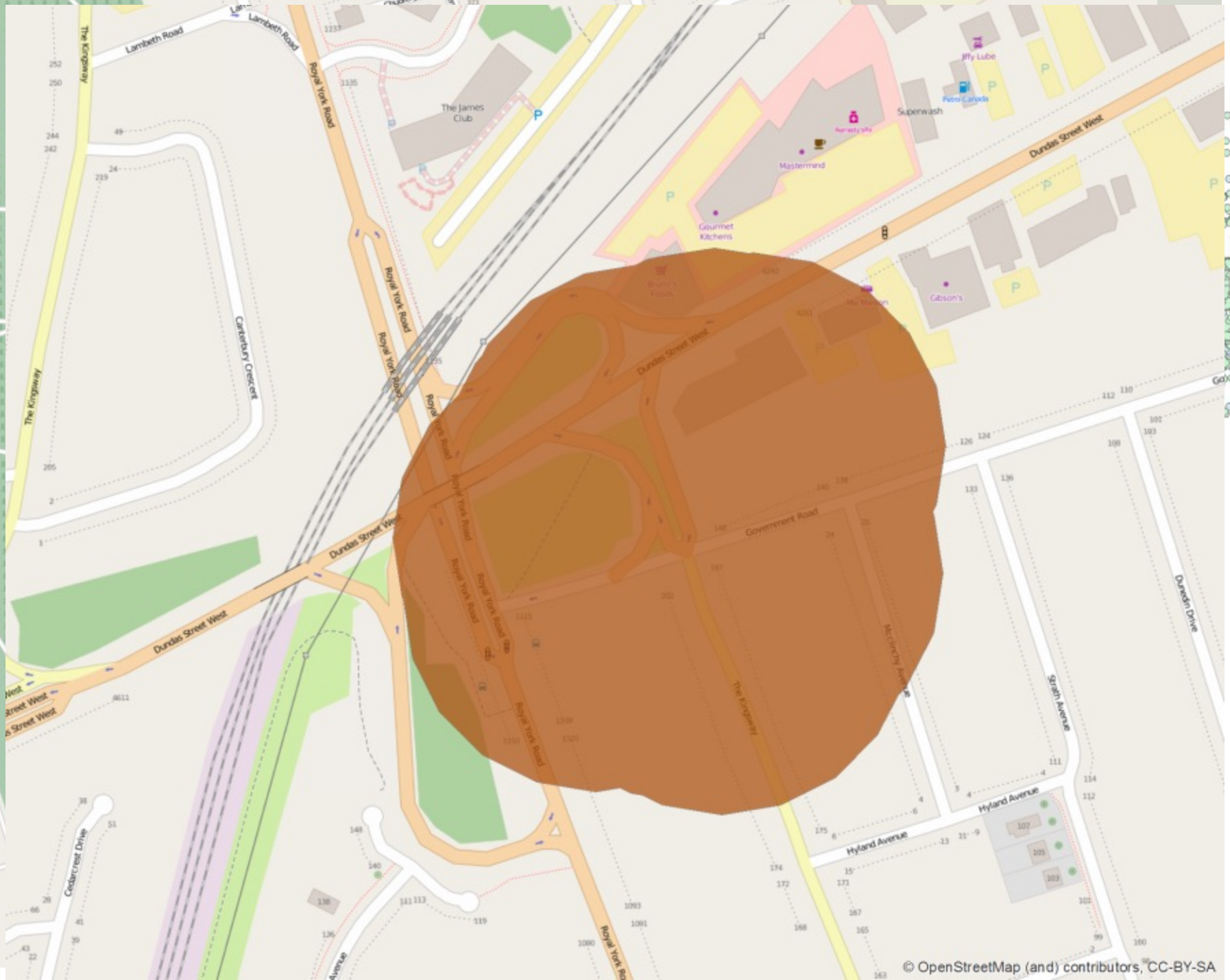
$$\mathbf{PAC_i} = \{PAC1_i, PAC2_i, PAC3_i\}$$


Art Gallery Of Ontario
Grange Park
Nathan Phillips Square
Eaton Centre
Moss Park
St James Park
St Lawrence Market
CBC
Convention Centre
Skydome
CN Tower
Air Canada Place
Harbour Square
Gardiner Expressway
Gardiner Expressway
Lake Ontario

Personal Activity Centers: Hypothesized Functions

- PAC1s
  - 1st Highest density
  - Concentrated spatial footprint
  - functions as: home?
- PAC2s
  - 2nd Highest density
  - Less concentrated spatial footprint
  - functions as: work?
- PAC3s
  - 3rd Highest Density
  - more dispersed
  - functions as: 'Digital Third Places' (e.g., centres of 'digital' social lives)




Lambeth Road
Royal York Road
The Kingsway
Canterbury Crescent
Dundas Street West
Government Road
Hyland Avenue
Strath Avenue
Dunedin Drive
Cedarcrest Drive
The James Club
Superwash
Jiffy Lube
Mastermind
Gourmet Kitchens
Gibson's
© OpenStreetMap (and) contributors, CC-BY-SA

Methods

1. Exploratory analysis of geosocial data spatial patterns
   - Summary statistics of key parameters
2. Test for differences in spatial patterns at the individual level over time
   - Wilcox tests: changes in size, change in density
3. Identify spatial patterns in space-use

[1] Robertson, C., Nelson, T.A., Boots, B. and Wulder, M.A. (2007) STAMP : Spatial – temporal analysis of moving polygons. Journal of Geographical Systems. 9: 207-227.

Case Study: Twitter in Toronto, ON

- 1.29 million tweets obtained for City of Toronto neighbourhoods in 2014 during summer and winter
- Generated individual PACs based on 100m threshold to identify potential 'places' of relevance to individuals
- Using individual patterns to explore place in geosocial data

What questions should we ask about this analysis?

- demographic biases
  - younger, technically competent
- location data?
  - geolocated tweets? → biased < 3 % of twitter users
    - income biases
  - user-tagged location → spatially biased towards significant places
    - tourist areas
  - location in profile? → may be incorrect, will not vary per user will be coarse
- biased toward positive sentiment


Here we go! #smwSocialEnt http://t.co/IvAimTt3yU
After dentist yogurt- ouch! :(#YOCOfy http://t.co/FI5uasqZ3j
Insight from the valley http://t.co/8STavKmXY5
Confessions http://t.co/ozcjiBLtdw
@thelizbuzz Agreed!
@jonkay And? Whom did you choose?
@Davejonesy Ditto.
Brunch (@ Windsor Arms Hotel) http://t.co/qo7zouzV6f
Don't leave without a hug @sarafalconer! #smwsocialgood
@juliebenoit h! Es-tu au @windsorarms ? #smwto
What a phenomenal talk by @DrBrynnWinegard at #SMWTO!
@DispersionPR
Mobile, mobile, mobile #SMWTO
@AARivard @LittleRoomInc exciting isn't it? @theofaktor
Quantified Self #
@DispersionPR how do I sign up for your promos list??
I'm disappointed. There's the clip. #topoli
My first #smwTO panel. #SMW
I NEED TO SEE IT.
.@LexPR @MargaretAtwood @smwto she's so lovely!
@katieedmonds1 awesome! I
ing to check out pocket. #SMWTO
@windsorarms thanks I appreciate the compliment.
agree 100%
Eating lunch at the Windor Arms Hotel. Whoo! :
@aimcook @SuperStarSaver great! Thanks and cheers!
@katieedmonds1 still around?
@Diana_Cowan deal!
@leighbryant #getthere
@Diana_Cowan meet after?
Great presentations@ #smwsocialgood today!
Oranje with 10 men without their number 10 #NEDvsARG
@FrauBeese just means we have a new place to check out :)
Is it ok to laugh? @benroy00 & @DarynJones don't seem to agree... #JFL42 #SMWTO http://t.co/Y2nx5Nm0fN
@NiamhTheWilson Do that during rush hour and watch yourself get beheaded. (If that is possible)
@ItsRowynne YES ROWYNNE YES !!!
This is good footy. ARG looking strong! #NEDvsARG
© OpenStreetMap (and) contributors, CC-BY-SA

Summary Statistics – highest order of PACs

Summer

AC1	AC2	AC3
60%	30%	9%

Winter

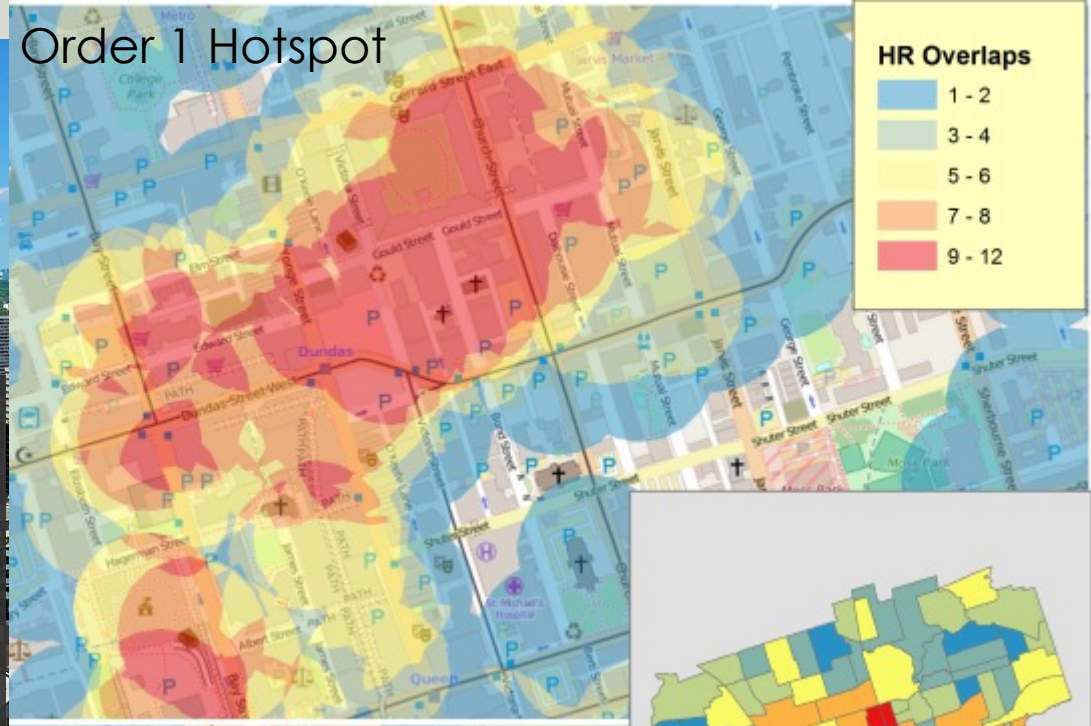
AC1	AC2	AC3
57%	33%	9%

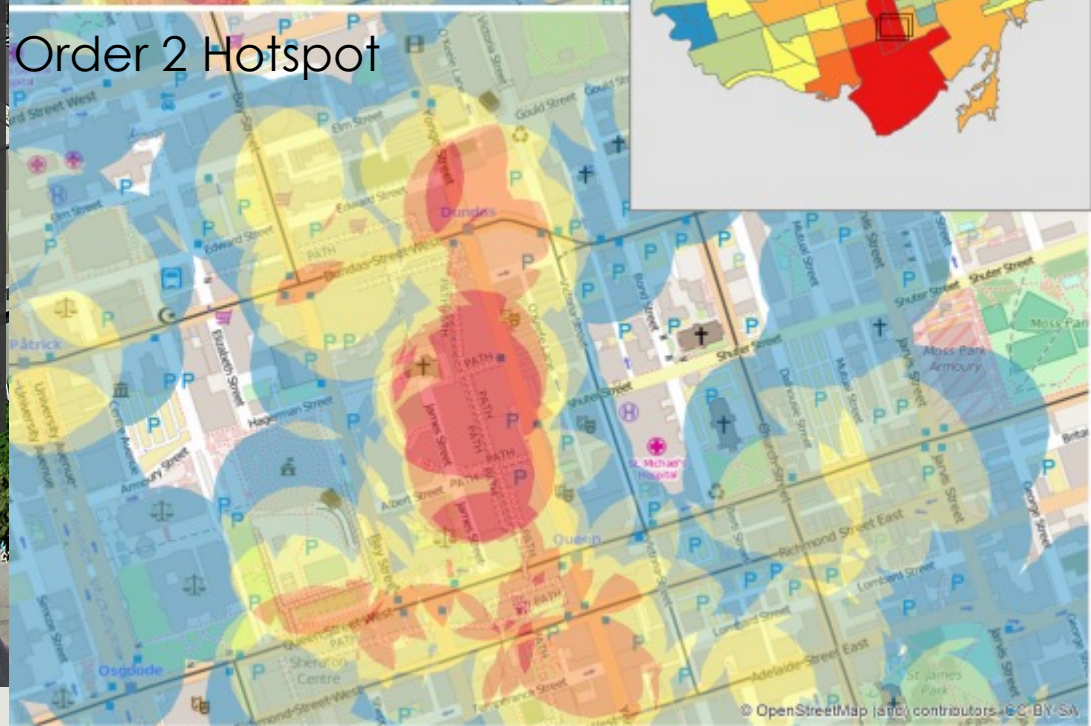
Summary Statistics

	Summer	Winter	
PAC Areas			
Avg area	7.68	7.57	W = 22085878, p-value = 0.09876
Avg area HR1 - Concurrent	8.04	8.22	V = 324510, p-value = 0.5362
Avg area HR2 - Concurrent	8.38	8.58	V = 12312, p-value = 0.6824
Avg area HR3 - Concurrent	9.08	8.41	V = 669, p-value = 0.1038
PAC Counts			
Tweets	60.05	62.61	W = 22350371, p-value = 0.00422
Tweets - Concurrent	82.80	102.64	V = 640085.5, p-value = 8.367e-13
Tweets HR1 - Concurrent	94.98	119.46	V = 398943.5, p-value = 2.171e-11
Tweets HR2 - Concurrent	43.23	45.22	V = 14203, p-value = 0.03027
Tweets HR3 - Concurrent	30.13	33.43	V = 1018, p-value = 0.2096

Place Identity?

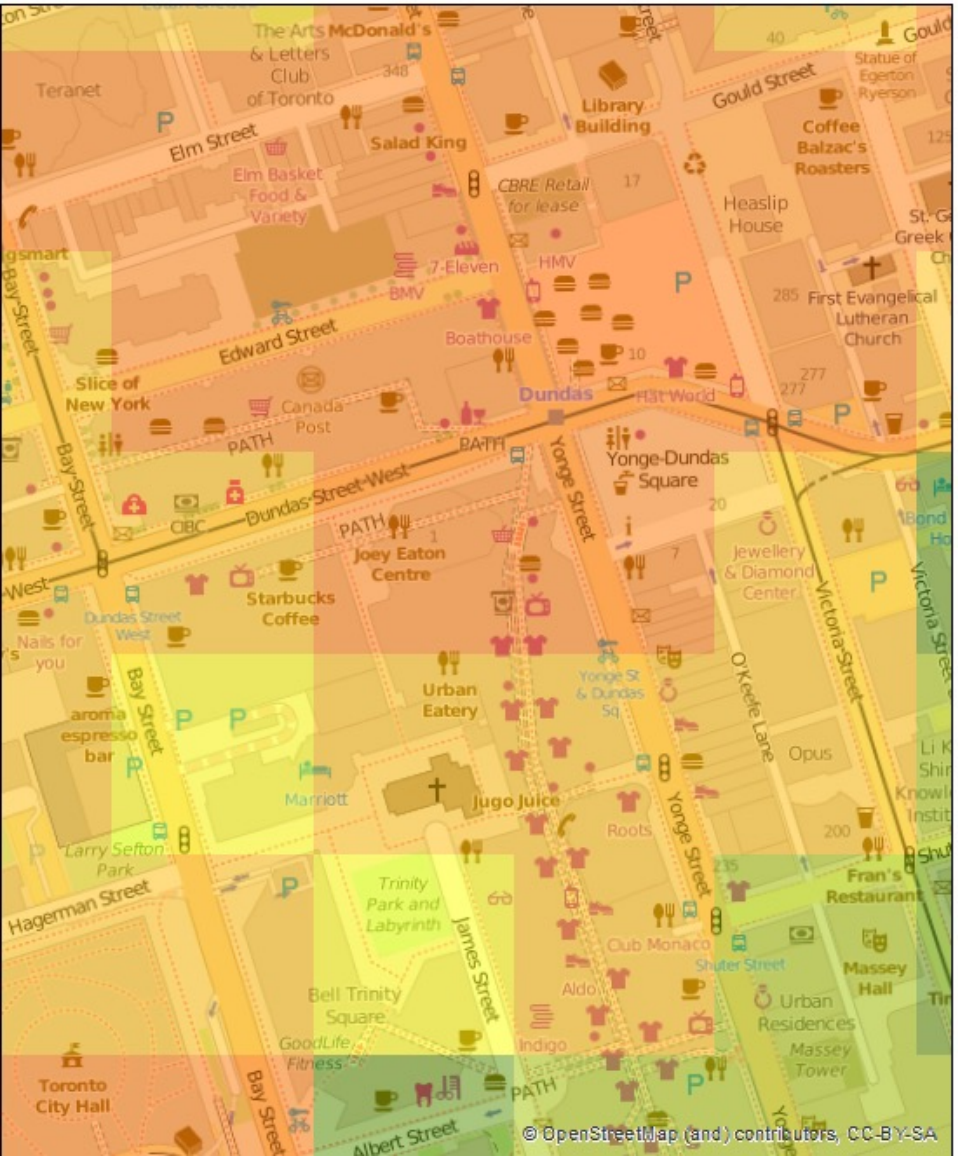
- Identifying overlaps of activity centres by order
  - Order 1 Hotspot
  - Order 2 Hotspot
- Clear functional aspects to these hotspots in terms of urban place



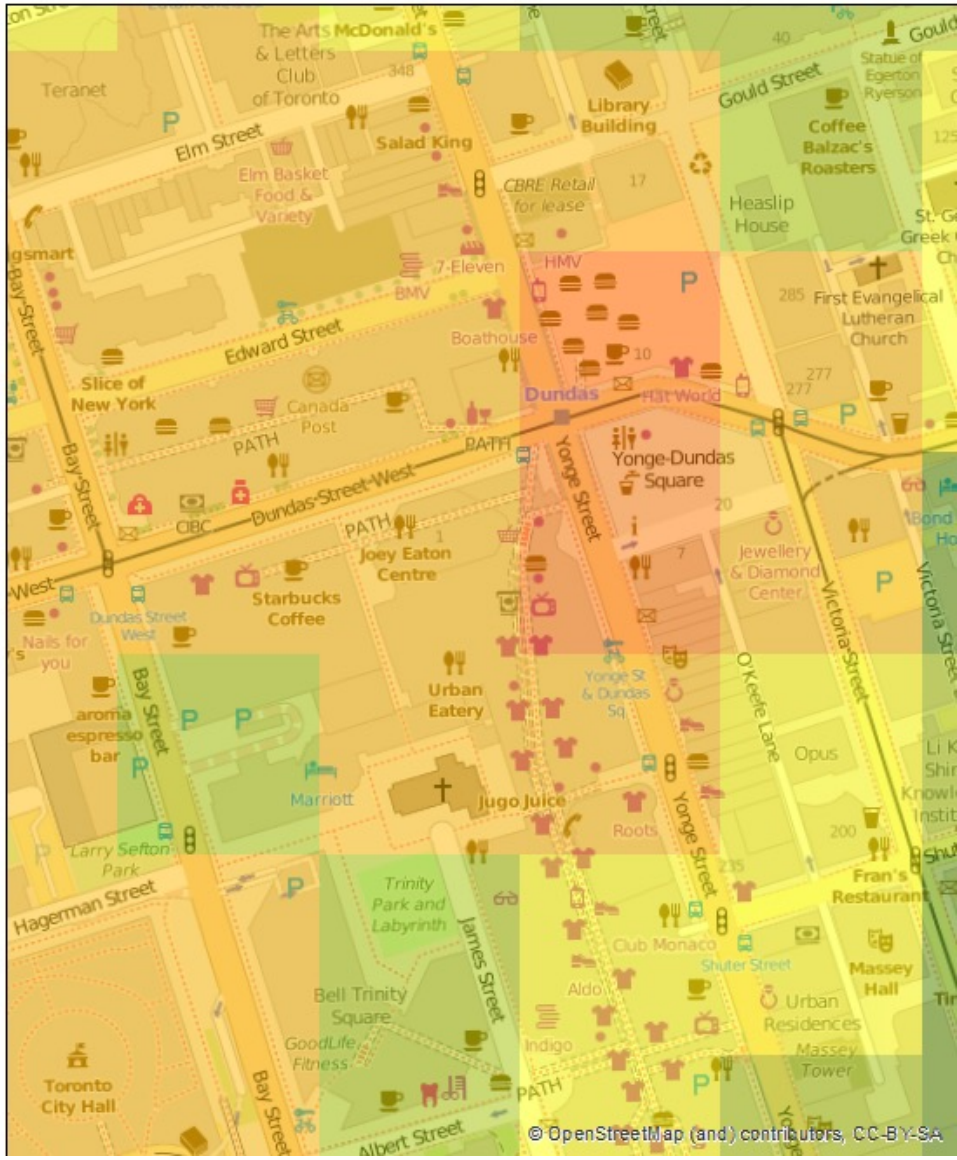


Order 1 PAC

Winter

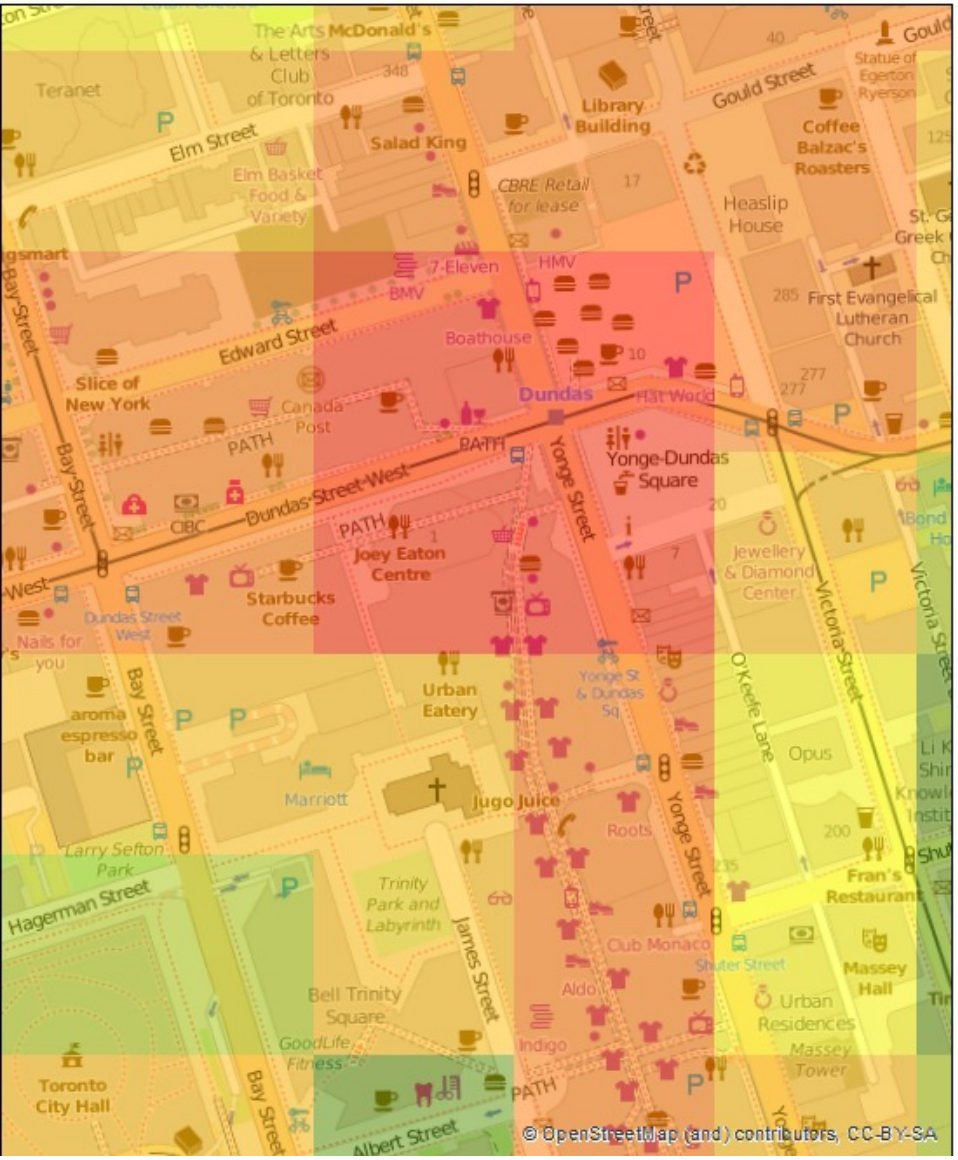


Summer

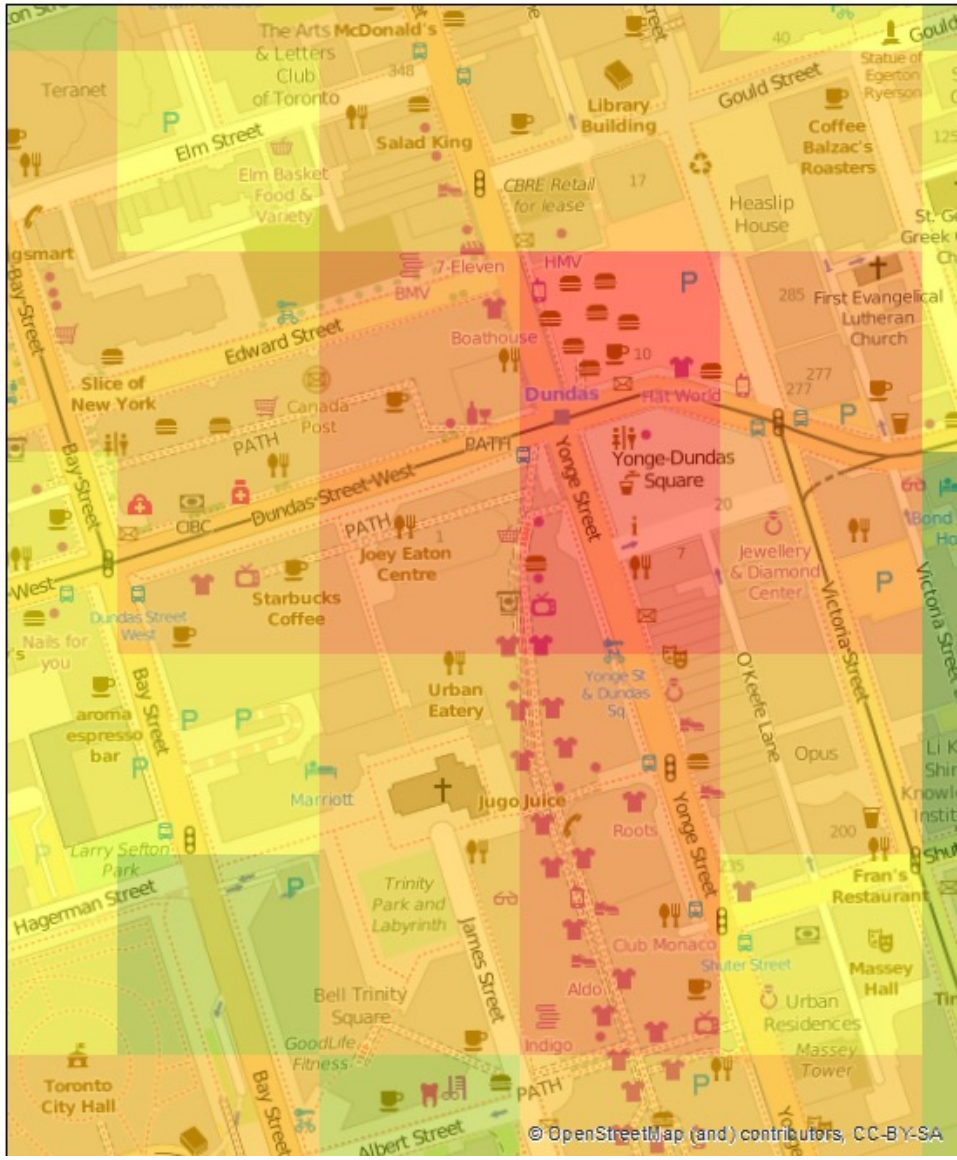


Order 2 PAC

Winter

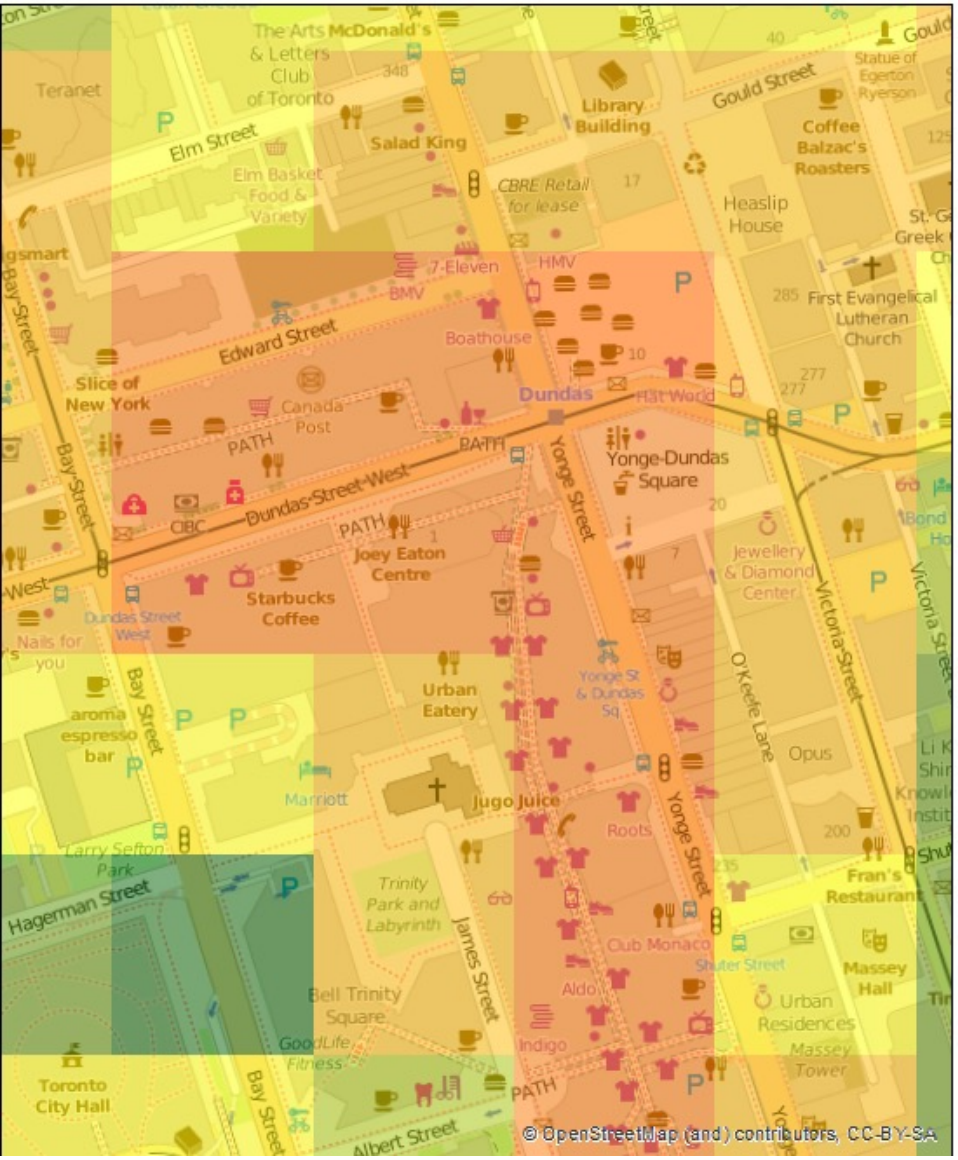


Summer

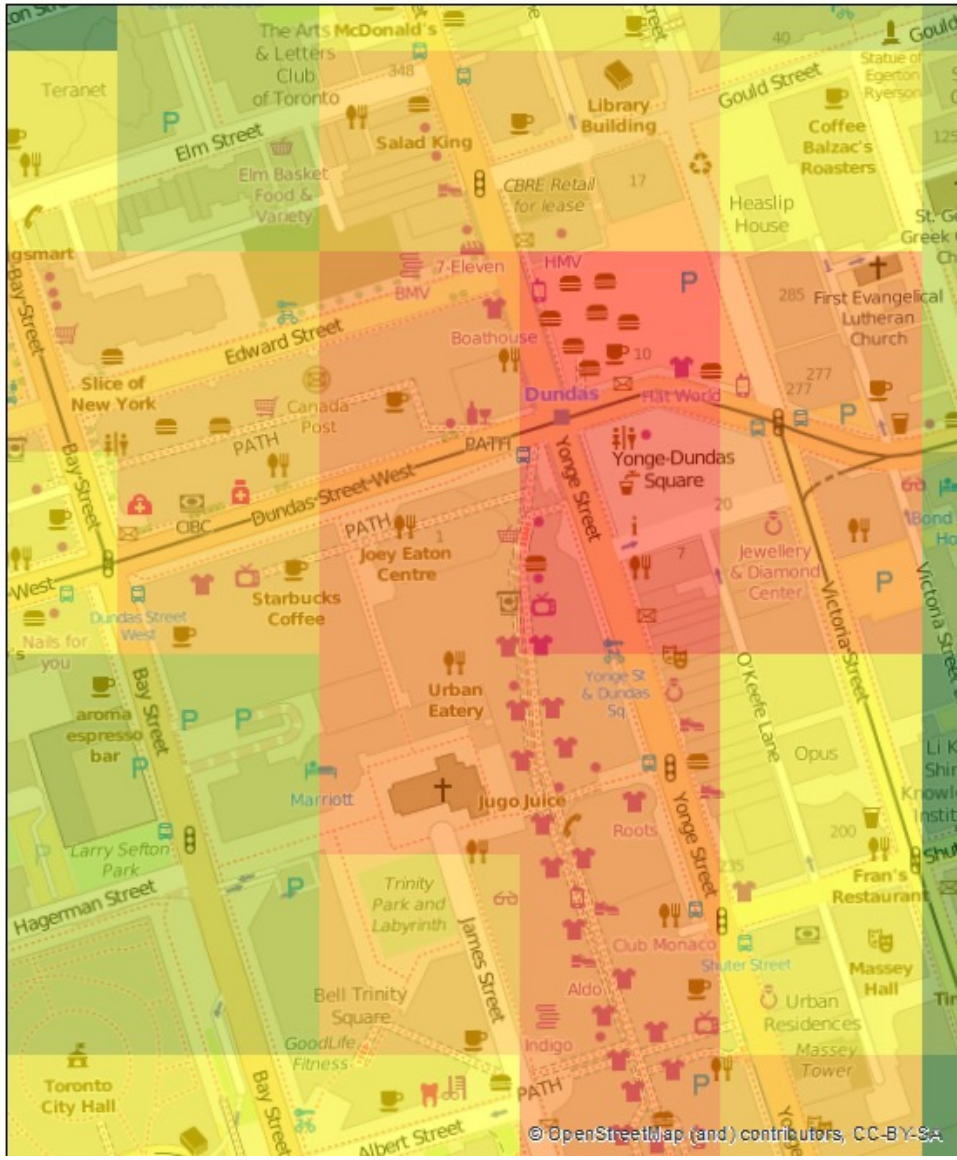


Order 3 PAC

Winter



Summer



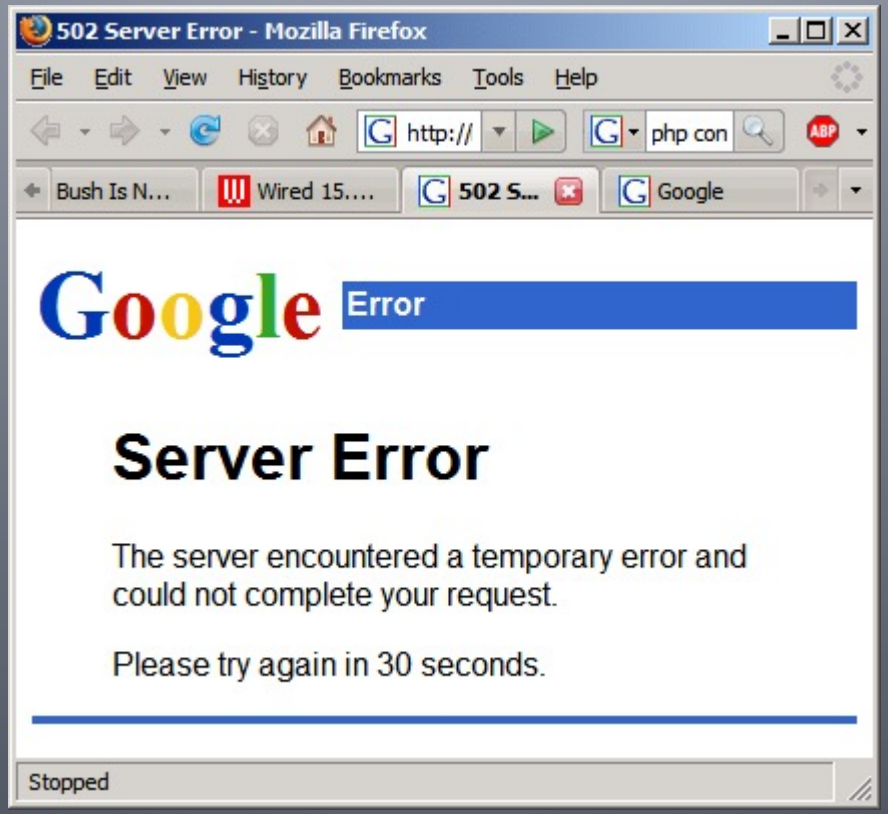
Summary of PAC mapping

- PAC1s more dispersed across the city
- Highest concentrations found for PAC3s
- Seasonal differences are evident
  - increasing Tweeting in Winter
  - more difference for PAC1 and PAC3 than PAC2

Test Critique

- Take 15 mins and scan the article
- Come up with one or two critiques
- https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(21)02796-3/fulltext#%20
- (hint china)

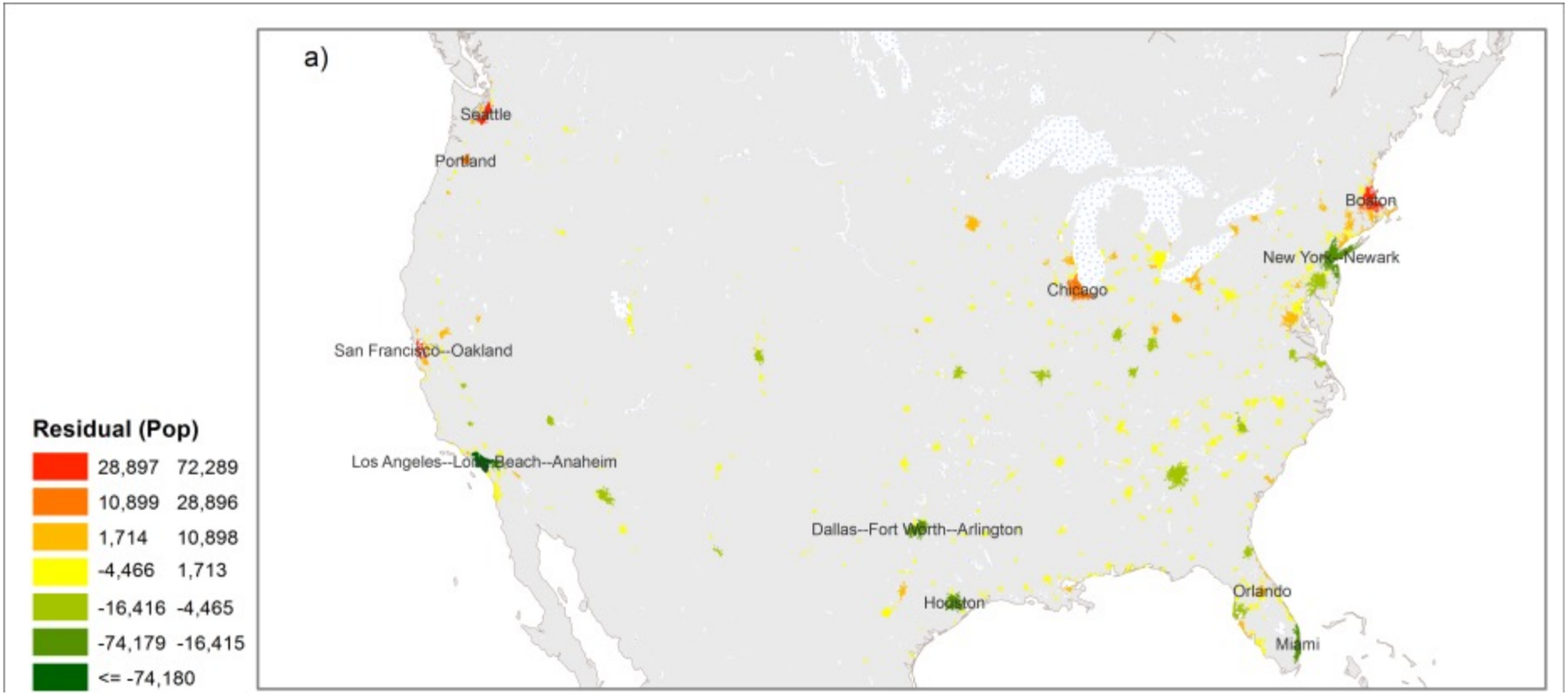
Critical Analysis of Big Data for Environment



We Need New Analysis Tools

- Visualization over statistical significance/models
- Patterns {space, time, theme} matter
  - Pattern changes matter more
  - Persistent, repeatable pattern changes … etc
- Opportunities outweigh challenges… geo+social analysis challenges similar to many big data problems…

Big Data Biases – Flickr data



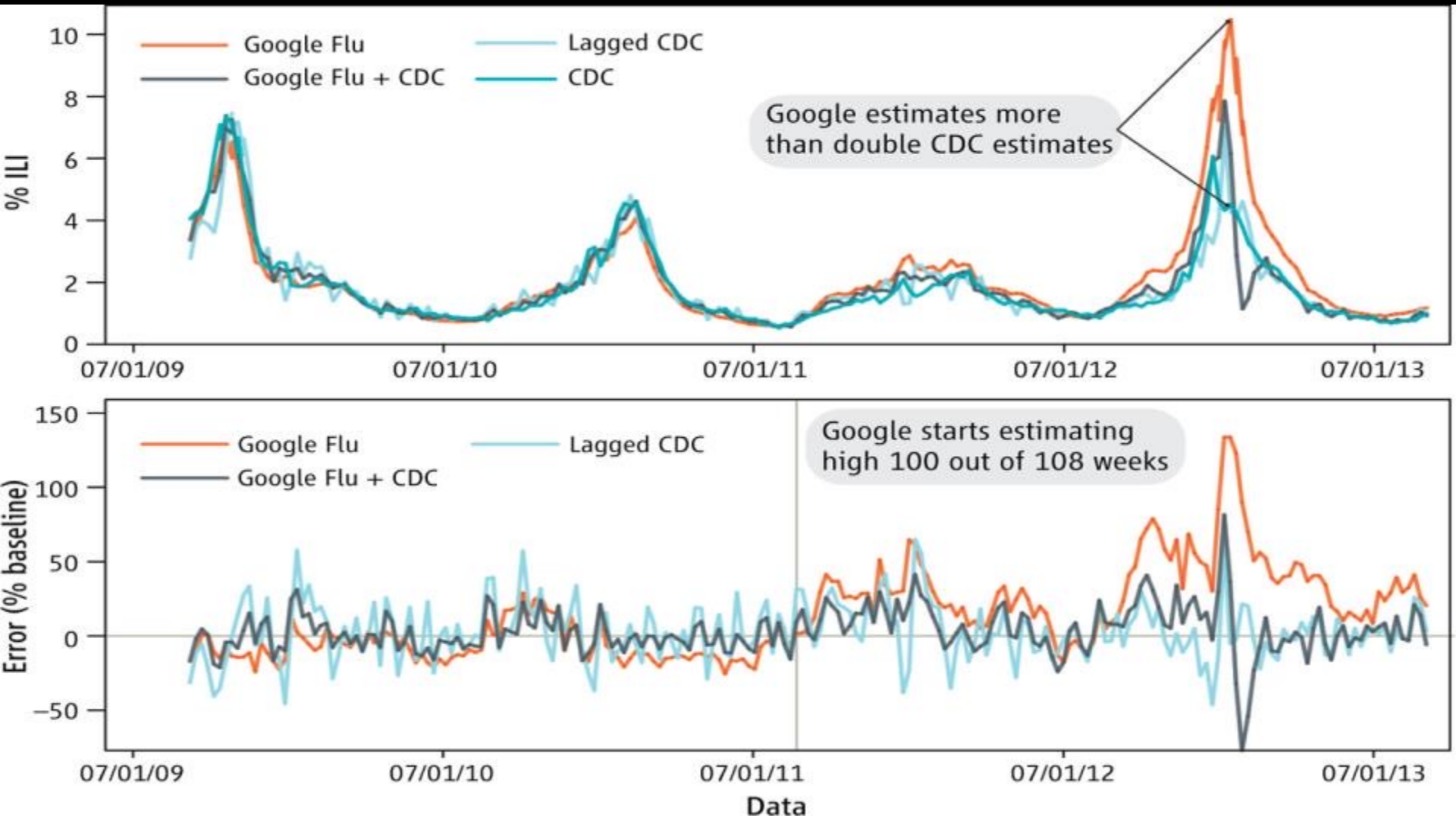
Over- and under-represented urban areas by urban area size based on geotagged photos across the United States.

Top Ranked Cities	Observed - Expected	Area (Ha)
New York--Newark, NY--NJ--CT	88056.7	889765.4
San Francisco--Oakland, CA	79122.4	130028.5
Seattle, WA	77479.3	263315.5
Chicago, IL--IN	56469.3	604148.5
Boston, MA--NH--RI	43927.3	474730.3
Bottom Ranked Cities		
Cincinnati, OH--KY--IN	-14016.7	194208.2
Charlotte, NC--SC	-15604.3	190272.3
Dallas--Fort Worth--Arlington, TX	-16057.0	447023.5
Houston, TX	-19421.7	420739.0
Atlanta, GA	-39757.0	658686.2

Poisson modeling results for selected covariates and geotagged photos at the urban area scale across the United States (* indicate significant at α = 0.05).

Coefficient	Estimate	p-value
Intercept	-3.572	<0.001*
GINI coefficient	**0.396**	**0.748**
% Under poverty line	**-0.113**	**<0.001***
% Unemployed	**0.046**	**0.010***
% Vacancy	0.009	0.117
% 1 unit detached housing	0.001	0.694
Estimated housing value	<0.001	0.436
% Walk to work	**0.097**	**<0.001***
Estimated travel time to work	**-0.031**	**<0.001***

Big Data, Big Mistakes?



Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. "The Parable of Google Flu: Traps in Big Data Analysis." Science 343 (6176) (March 14): 1203–

What are the challenges of using Big Data for environmental research?

1. Individuals vs general patterns – context matters?
2. Who is included and who is excluded? What are the dimensions of the digital divide?
3. How is data consumed by algorithms, what biases are embedded?
4. What locations are covered, how do they reflect our values and view of the world?
5. Does big data promote democracy and empower disadvantaged groups or serve as a tool to further marginalize certain groups?
6. Data quality – does it matter?