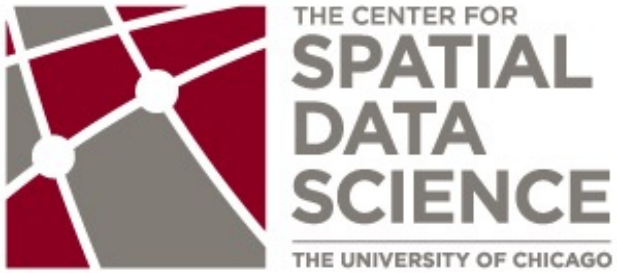
Spatial Data, Spatial Analysis and Spatial Data Science

Luc Anselin



http://spatial.uchicago.edu

spatial thinking in the social sciences

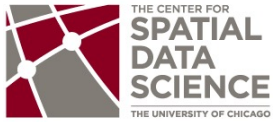
spatial analysis

spatial data science

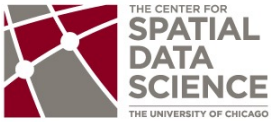
spatial data types and research questions

pitfalls

examples



Spatial Thinking in the Social Sciences



- Motivation - Substantive

  from atomistic decision units to social-spatial interaction

    peer-effects, copy-catting, diffusion

    spatial imprint of social networks/interaction

  spatial externalities

    spatial spillovers, spatial multipliers

- Motivation - Substantive (continued)

  "social facts are located" (Abbott 1997)

    spatial mismatch

    spatial disparities

  spatial context

    neighborhood effects

- Motivation - Data

  geo-located observations

    street addresses of crimes, sensor data,
    social media data

  mismatch between the spatial scale of the process and the spatial scale of the observations

    administrative units (e.g., census tracts) are not behavioral units (e.g., neighborhoods)

- Motivation - Data (continued)
  - error terms show systematic patterns
    - neighborhood effects in individual house price models
    - distance decay (precision decreases with distance from sensors)
  - change of (spatial) support problem
    - data at different spatial scales, nested or overlapping
      - census block groups into census tracts
      - school districts and census tracts

Some Examples

Of Time and Space: The Contemporary Relevance of the Chicago School*

ANDREW ABBOTT, *University of Chicago*

Abstract

This essay argues that sociology's major current problems are intellectual. It traces these problems to the exhaustion of the current "variables paradigm" and considers the Chicago School's "contextualist paradigm" as an alternative. Examples of new methodologies founded on contextual thinking are considered.

Anniversaries are often valedictions. A centennial sometimes shows an association to be moribund, just as a diamond jubilee may reveal a queen's irrelevance and a golden anniversary finds many a marriage dead. By contrast, living social relations celebrate themselves daily. Anniversaries merely punctuate their excitement.

What then are we to make of this centennial year of sociology at the University of Chicago? Is it simply a time for eulogy? After all, Chicago dominance of sociology is half a century gone. And while the Chicago tradition renewed itself after the war in Goffman, Becker, Janowitz, and their like, many of Chicago's most distinguished alumni since its dominant years belong more to the mainstream than to the Chicago tradition proper: methodologists like Stouffer and Duncan, demographers like Hauser and Keyfitz, macrosociologists like Bendix and Wilensky. Nonetheless, at the heart of the Chicago tradition lie insights central to the advancement of contemporary sociology. Therefore, I do not today eulogize the Chicago tradition. One eulogizes only the dead.[1]

** This article sparked a lot of commentary. Surprisingly, helpful comments came not only from people I knew well, but also from relative strangers. I have therefore had more help with this article than with virtually anything else I have written. The following all contributed substantial comments: Rebecca Adams, Joan Aldous, Margo Anderson, James Coleman, Claude Fischer, Jeffrey Goldfarb, David Maines, Donald Levine, Douglas Mitchell, John Modell, John Padgett, Moishe Postone, and Charles Tilly. I would like to dedicate this essay to the memory of Morris Janowitz, who taught me and many others about the Chicago School. Address correspondence to Andrew Abbott, Department of Sociology, 1126 East 59th St., University of Chicago, Chicago, IL 60637.*

 Social Forces, June 1997, 75(4):1149-82

Abbott (1997) Chicago School

INTERNATIONAL REGIONAL SCIENCE REVIEW 23, 2: 139–159 (April 2000)

TOWARD SPATIALLY INTEGRATED SOCIAL SCIENCE

MICHAEL F. GOODCHILD
Department of Geography, University of California, Santa Barbara, good@ncgia.ucsb.edu

LUC ANSELIN
Department of Agricultural and Consumer Economics, University of Illinois at Urbana–Champaign, anselin@uiuc.edu

RICHARD P. APPELBAUM
Department of Sociology, University of California, Santa Barbara, appelbau@alishaw.sscf.ucsb.edu

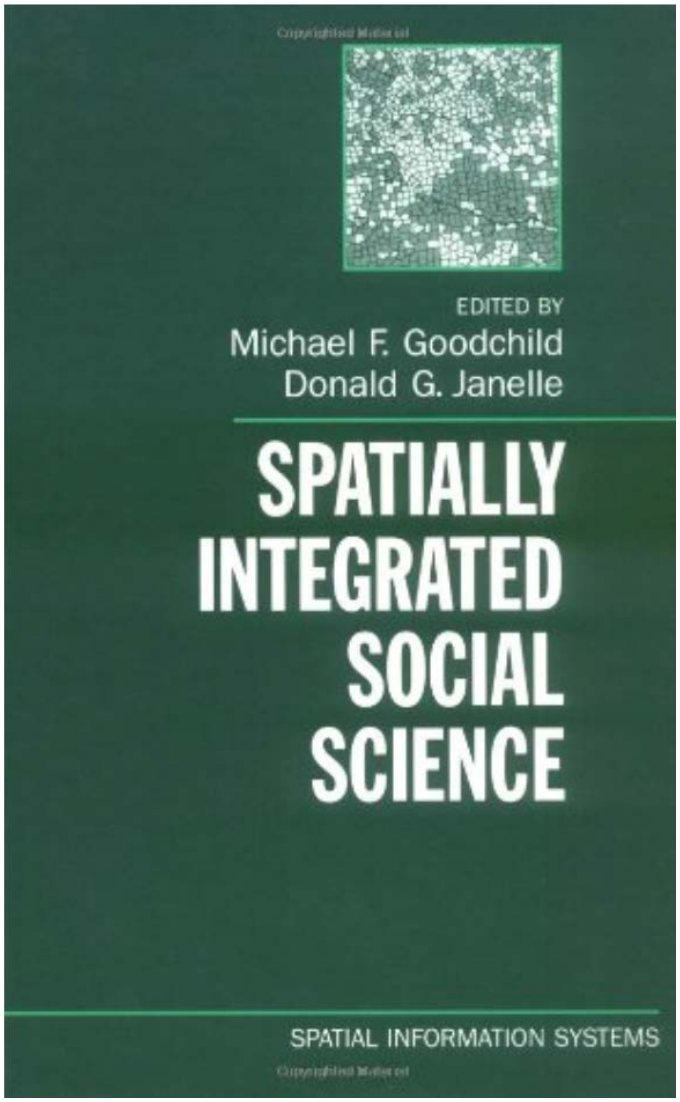
BARBARA HERR HARTHORN
Department of Anthropology, University of California, Santa Barbara, bharthor@omni.ucsb.edu

This article outlines the motivation for a spatial approach as a novel focus for cross-disciplinary interaction and research in the social and behavioral sciences. The authors review the emerging interest in space and place in the recent social science literature and develop a vision for a spatially integrated social science. This vision provides the conceptual basis for a program of six activities designed to promote a spatial perspective: learning resources, workshops, best-practice examples, place-based search, software tools, and a virtual community. The six programs will be informed by advances in the methods, technologies, and principles underlying spatial information science.

The analysis of space and place has become an increasingly pivotal component of social science research in the past two decades. In part, this can be attributed to the transformation of social space around the globe, accompanied by shifts of varying degrees of magnitude in social science conceptualizing and theorizing. One aspect of these changes is subsumed under the general notion of "space-time compres-

---

This article is a revised and shortened version of a proposal to the U.S. National Science Foundation (NSF) titled "SPESS: A Center for Spatially Enabled Social Science," which resulted in a five-year award to the University of California, Santa Barbara. Earlier versions were presented at the Interuniversity Consortium for Political and Social Research Meeting of Official Representatives on "Approaching



spatially integrated social science
Goodchild et al (2000), Goodchild and Janelle (2004)

across a rich vari-
when it breached
ame embedded in

terramechanics
rk by Bekker (*2*).
ocomotion began
vehicle power per
comotion speed.
d on the dimen-
, is shown in the
original figure by
cludes the Gabri-
ich was first pre-
ical fit of the best
r a range of vehi-
) and speeds (*10*).
elli–von Karman
ded animals and
size that vehicles
nds and soils have
hen operating on

ata for pedestri-
es from Gabrielli
updated by Yong
e figure presents
ecific mechani-
ard (*5*); the mea-
exopod robots on
surfaces (*4*); and
RV and the MER

MEDICINE

Spatial Turn in Health Research

Douglas B. Richardson,[1] Nora D. Volkow,[2] Mei-Po Kwan,[3] Robert M. Kaplan,[4] Michael F. Goodchild,[5] Robert T. Croyle[6]

Developments in geographic science and technology can increase our understanding of disease prevalence, etiology, transmission, and treatment.

Spatial analysis using maps to associate geographic information with disease can be traced as far back as the 17th century. Today, recent developments and the widespread diffusion of geospatial data acquisition technologies are enabling creation of highly accurate spatial (and temporal) data relevant to health research. This has the potential to increase our understanding of the prevalence, etiology, transmission, and treatment of many diseases.

New approaches in geography and related fields, capitalizing on advances in technologies such as geographic information systems (GIS), the Global Positioning System (GPS), satellite remote sensing, and computer cartography, are often referred to collectively as geographic information science (*1*, *2*). GPS and related systems make it possible to integrate highly accurate geographic location and time with virtually any observation. GIS provides the means to store, share, analyze, and visualize real-time and archived spatial data. It also permits the integration of multiple layers of interdisciplinary spatial data, such as health, environmental, genomic, social, or demo-

[1]Association of American Geographers (AAG), 1710 16th Street, NW, Washington, DC 20009, USA. [2]National Institute on Drug Abuse, National Institutes of Health, Bethesda, MD 20852, USA. [3]Department of Geography and Geographic Information Science, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA. [4]Office of Behavioral and Social Sciences Research, National Institutes of Health, Bethesda, MD 20892, USA. [5]Department of Geography, University of California, Santa Barbara, CA 93106, USA. [6]Division of Cancer Control and Population Sciences, National Cancer Institute, Rockville, MD 20852, USA. E-mail: drichardson@aag.org

22 MARCH 2013 VOL 339 **SCIENCE** www.sciencemag.org
Published by AAAS

spatial turn in
health research
Richardson et al (2013)

THE WHITE HOUSE
WASHINGTON

August 11, 2009

M-09-28

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Peter R. Orszag, Office of Management and Budget
Melody Barnes, Domestic Policy Council
Adolfo Carrion, Office of Urban Affairs
Lawrence Summers, National Economic Council

SUBJECT: Developing Effective Place-Based Policies for the FY 2011 Budget

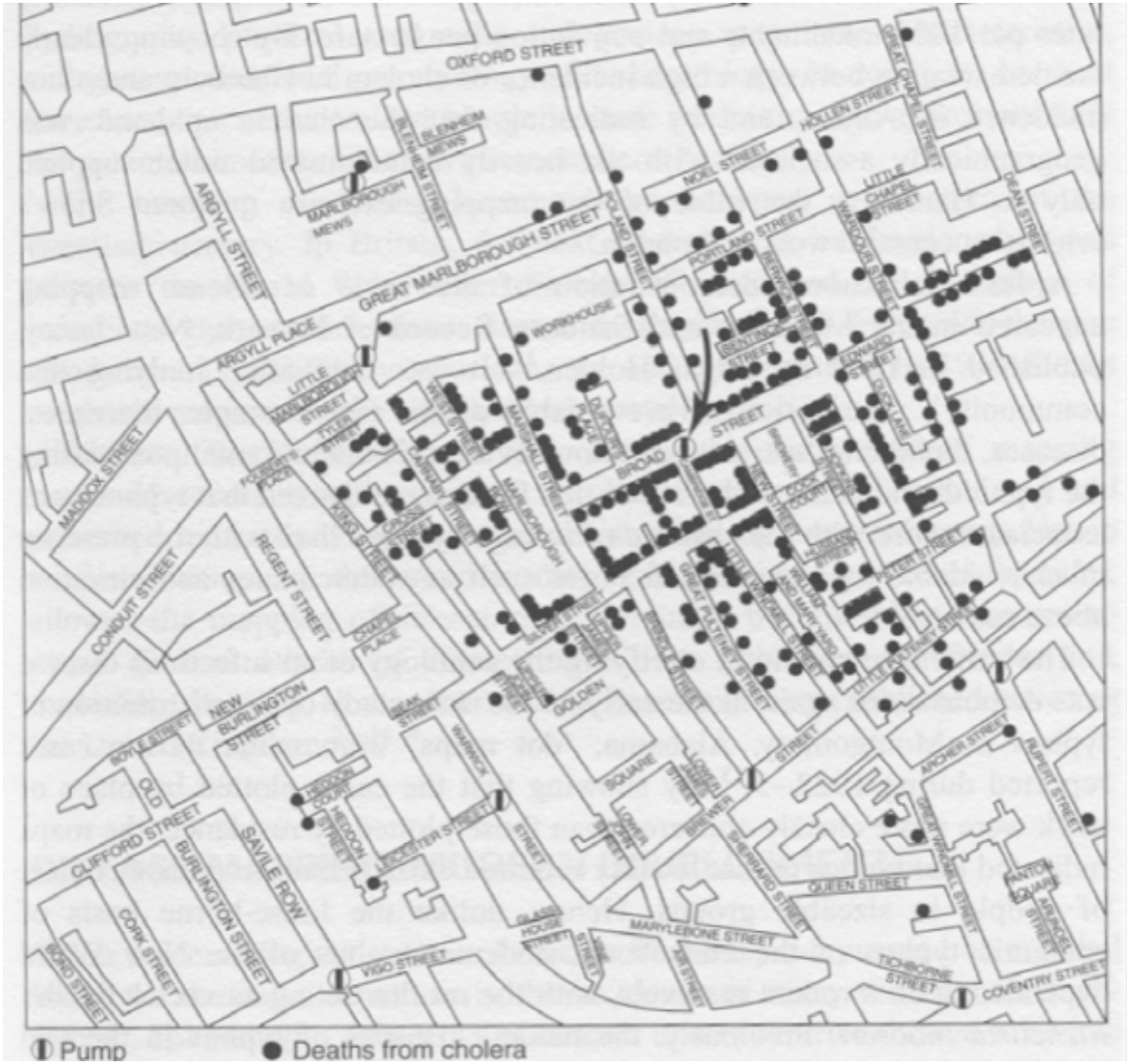
This guidance memorandum outlines policy principles meant to advance the Administration's domestic and fiscal priorities and to increase the impact of government dollars by leveraging place-conscious planning and place-based programming.

The guidance outlined here is preliminary. It supports an important interagency process focused on investing in what works by evaluating existing place-based policies and identifying potential reforms and areas for interagency coordination. Our immediate objective is to develop proposals for the FY2011 Budget that advance this Administration's policy priorities in the most effective ways whether by improving place-based strategies already operating or by adopting such strategies where there is significant potential for impact on a problem(s).

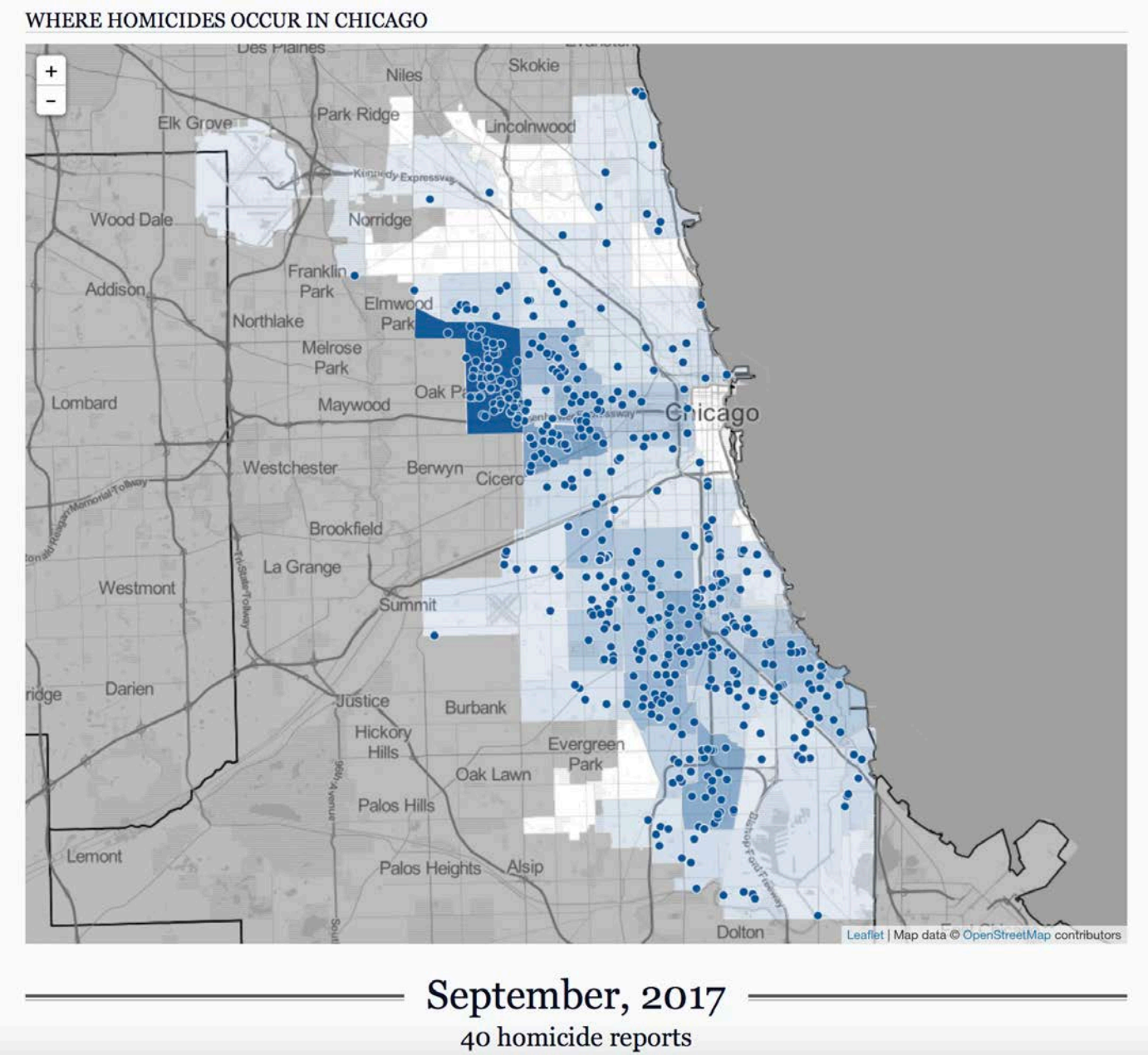
Place-based policies leverage investments by focusing resources in targeted places and drawing on the compounding effect of well-coordinated action. Effective place-based policies can influence how rural and metropolitan areas develop, how well they function as places to live, work, operate a business, preserve heritage, and more. Such policies can also streamline otherwise redundant and disconnected programs.

OMB Circular M-09-28
Effective Place-Based Policies

Spatial Analysis



Dr. John Snow Map of 1854 London Cholera



Modern-Day Version (Chicago Tribune)

- What is Spatial Analysis

  beyond mapping

  - added value

    transformations, manipulations and application of analytical methods to spatial (geographic) data
    (Goodchild et al, Geospatial Analysis)

  (geospatial) knowledge discovery

    specialized form of KDD,
    knowledge discovery from data(bases)

  - from data to information to knowledge to wisdom

- Spatial Analytics Questions

  where do things happen: patterns, clusters, hot spots, disparities

  why do they happen where they happen: location decisions

  how does where things happen affect other things (context, environment) and how does context affect what happens: interaction

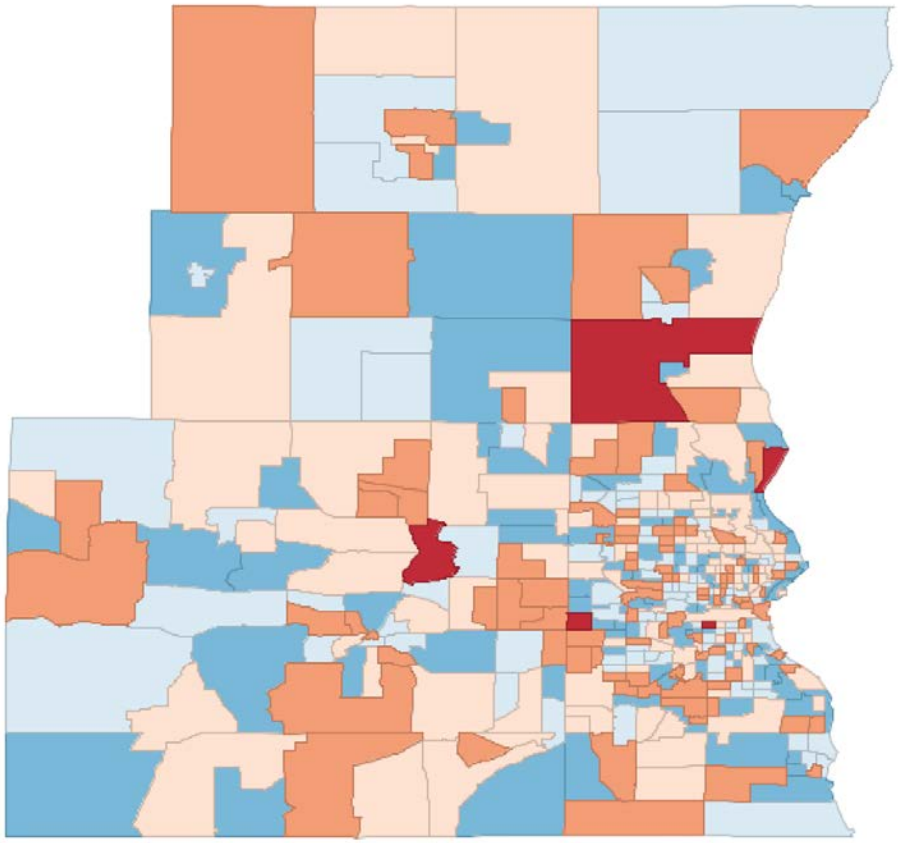
  where should things be located: optimization

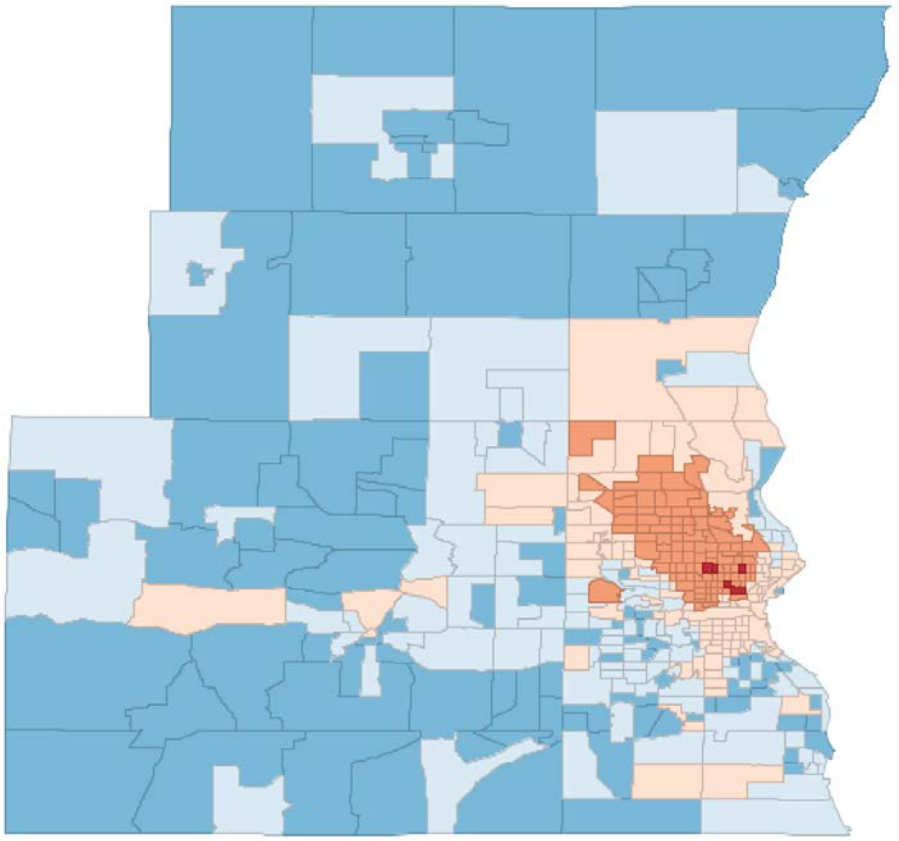
- When is Analysis Spatial?

  geo-spatial data:
  location + value (attribute)

  "non-spatial" analysis:
  location does NOT matter = locational invariance

  spatial analysis:
  when the location changes, the information content of the data changes

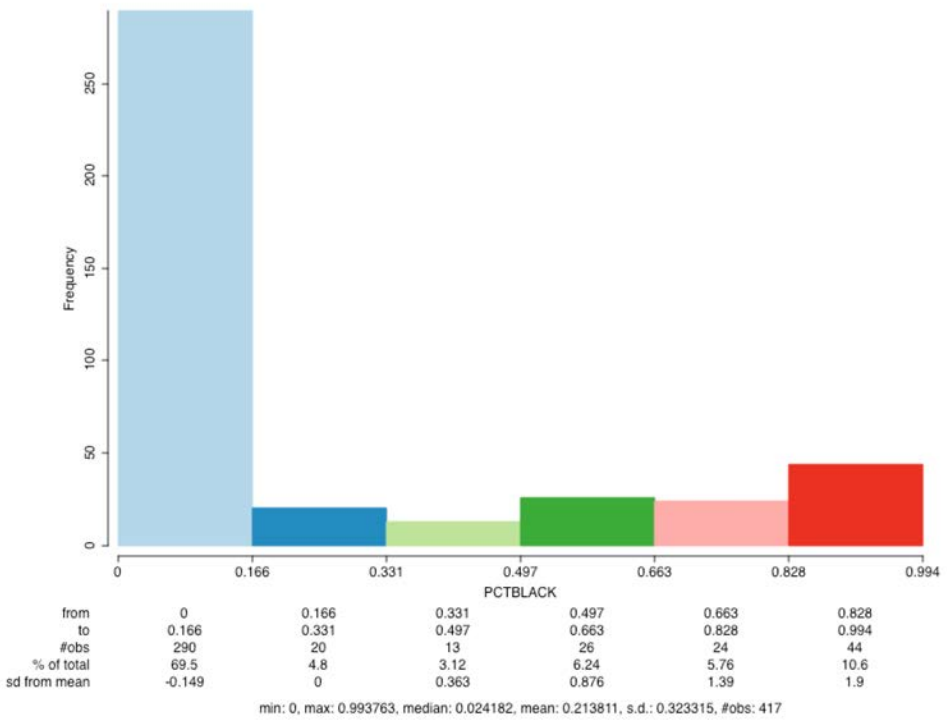




Hinge=1.5: PCTBLACK
- Lower outlier (0)
- < 25% (104)
- 25% - 50% (104)
- 50% - 75% (105)
- > 75% (99)
- Upper outlier (5)

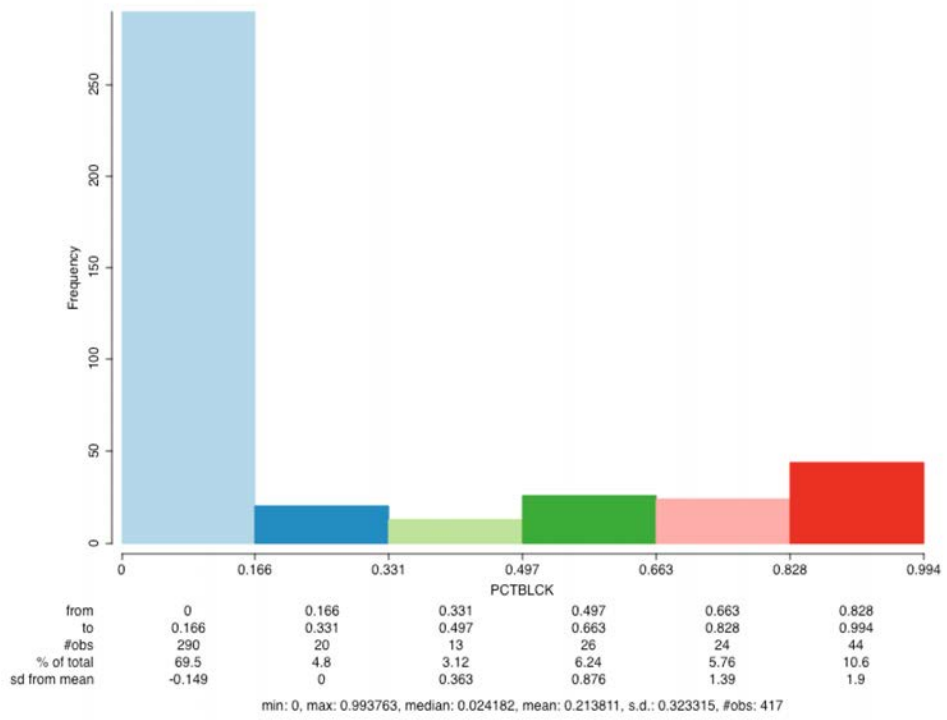
Hinge=1.5: PCTBLCK
- Lower outlier (0)
- < 25% (104)
- 25% - 50% (104)
- 50% - 75% (105)
- > 75% (99)
- Upper outlier (5)

Spatial Distribution



from	0	0.166	0.331	0.497	0.663	0.828
to	0.166	0.331	0.497	0.663	0.828	0.994
#obs	290	20	13	26	24	44
% of total	69.5	4.8	3.12	6.24	5.76	10.6
sd from mean	-0.149	0	0.363	0.876	1.39	1.9

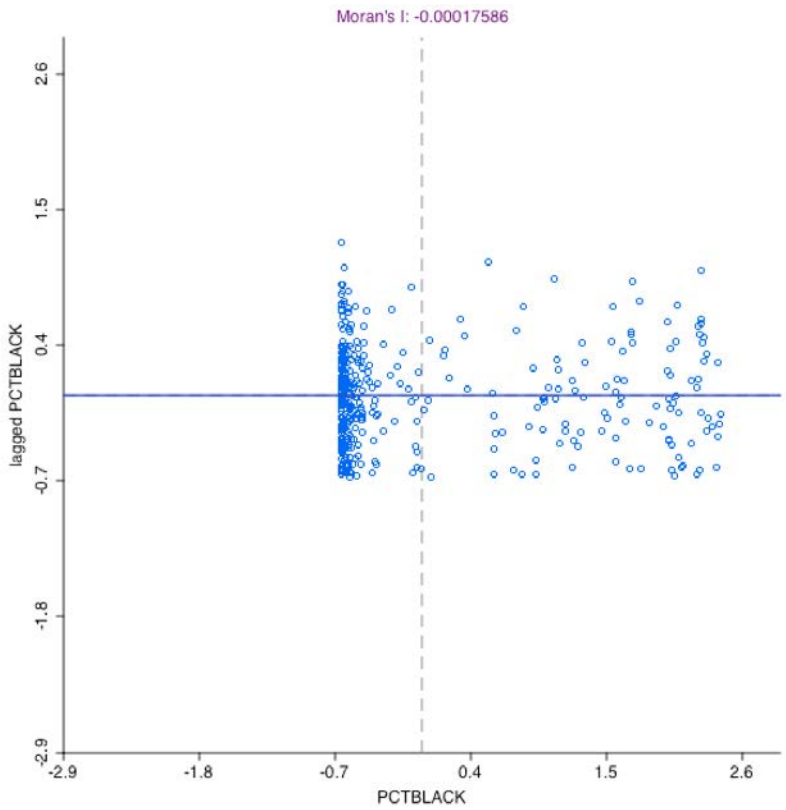
min: 0, max: 0.993763, median: 0.024182, mean: 0.213811, s.d.: 0.323315, #obs: 417

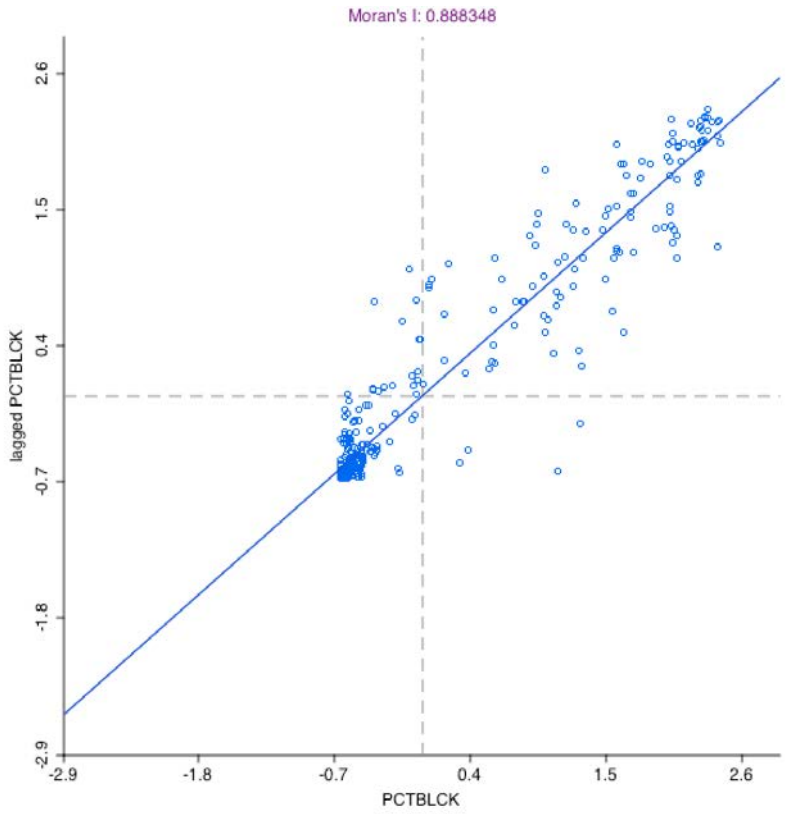


from	0	0.166	0.331	0.497	0.663	0.828
to	0.166	0.331	0.497	0.663	0.828	0.994
#obs	290	20	13	26	24	44
% of total	69.5	4.8	3.12	6.24	5.76	10.6
sd from mean	-0.149	0	0.363	0.876	1.39	1.9

min: 0, max: 0.993763, median: 0.024182, mean: 0.213811, s.d.: 0.323315, #obs: 417

A-Spatial Distribution
Histogram





Spatial Analysis
Global Spatial Autocorrelation
Moran Scatter Plot

- Components of Spatial Data Analytics

  mapping and geovisualization

    showing interesting patterns

  exploratory spatial data analysis

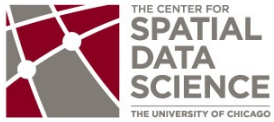
    discovering interesting patterns

  spatial modeling

    explaining interesting patterns

    optimization, simulation, prediction

Spatial Data Science



Big Data

- The Big Data Phenomenon

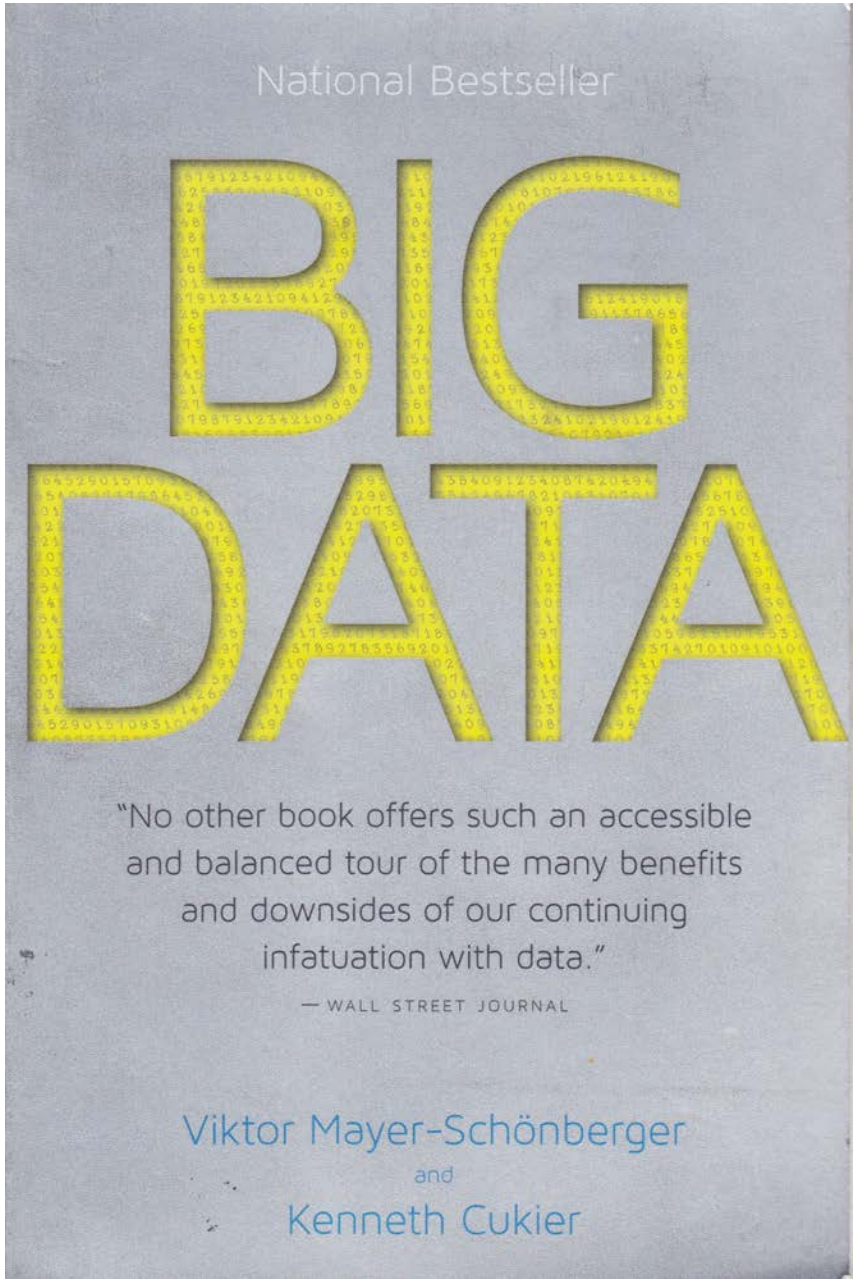
  ill-defined, you know it when you see it

  cannot be handled with current x

    x = hardware, memory, software, methodology, etc.

  the five "v"

  - volume, velocity, variety, value, veracity



- Big Data Issues

  sample size = population or is N = 1

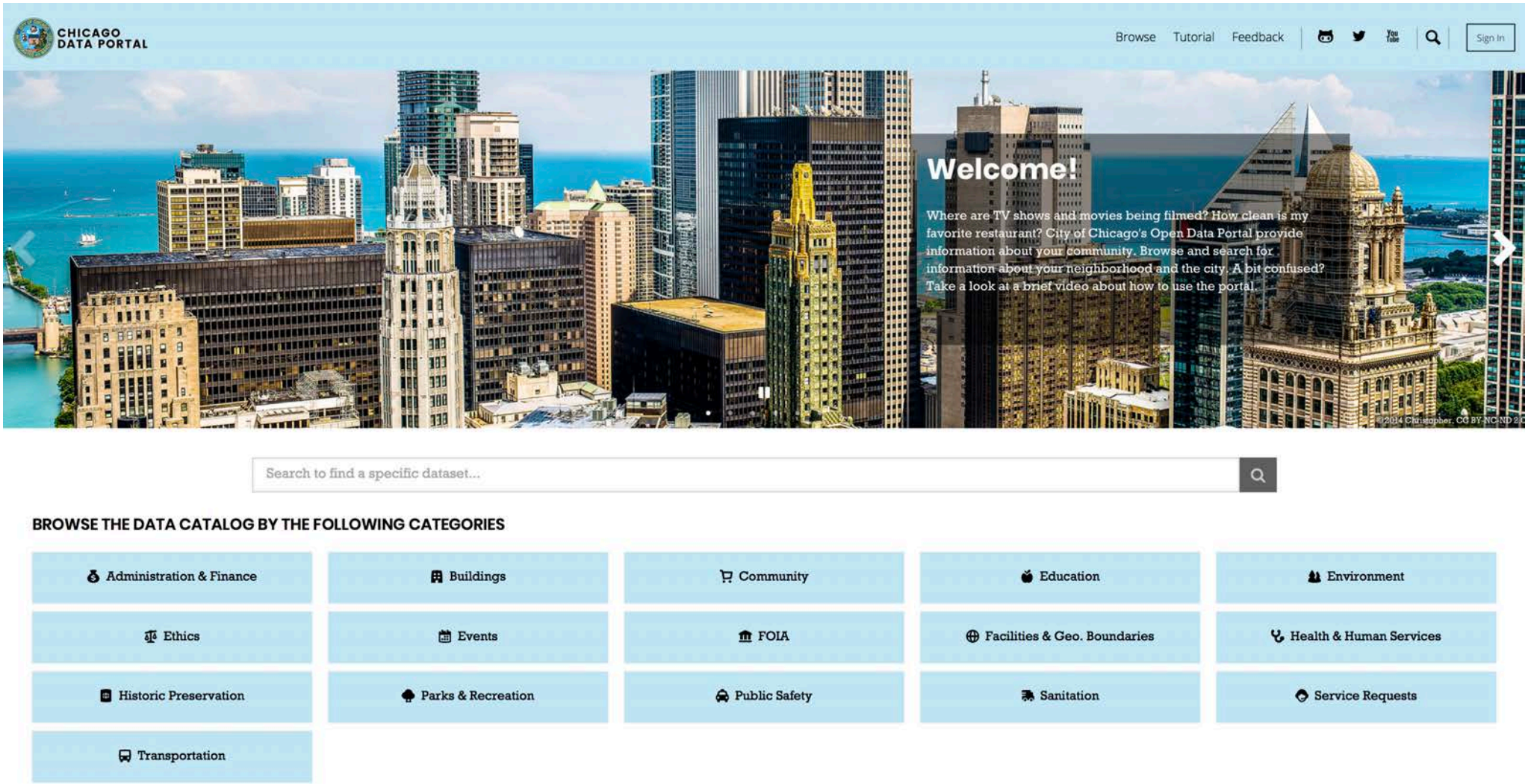
  size of data set compensates for imprecision, lack of sampling framework, measurement error, etc. , or does it?

  correlation, not causation

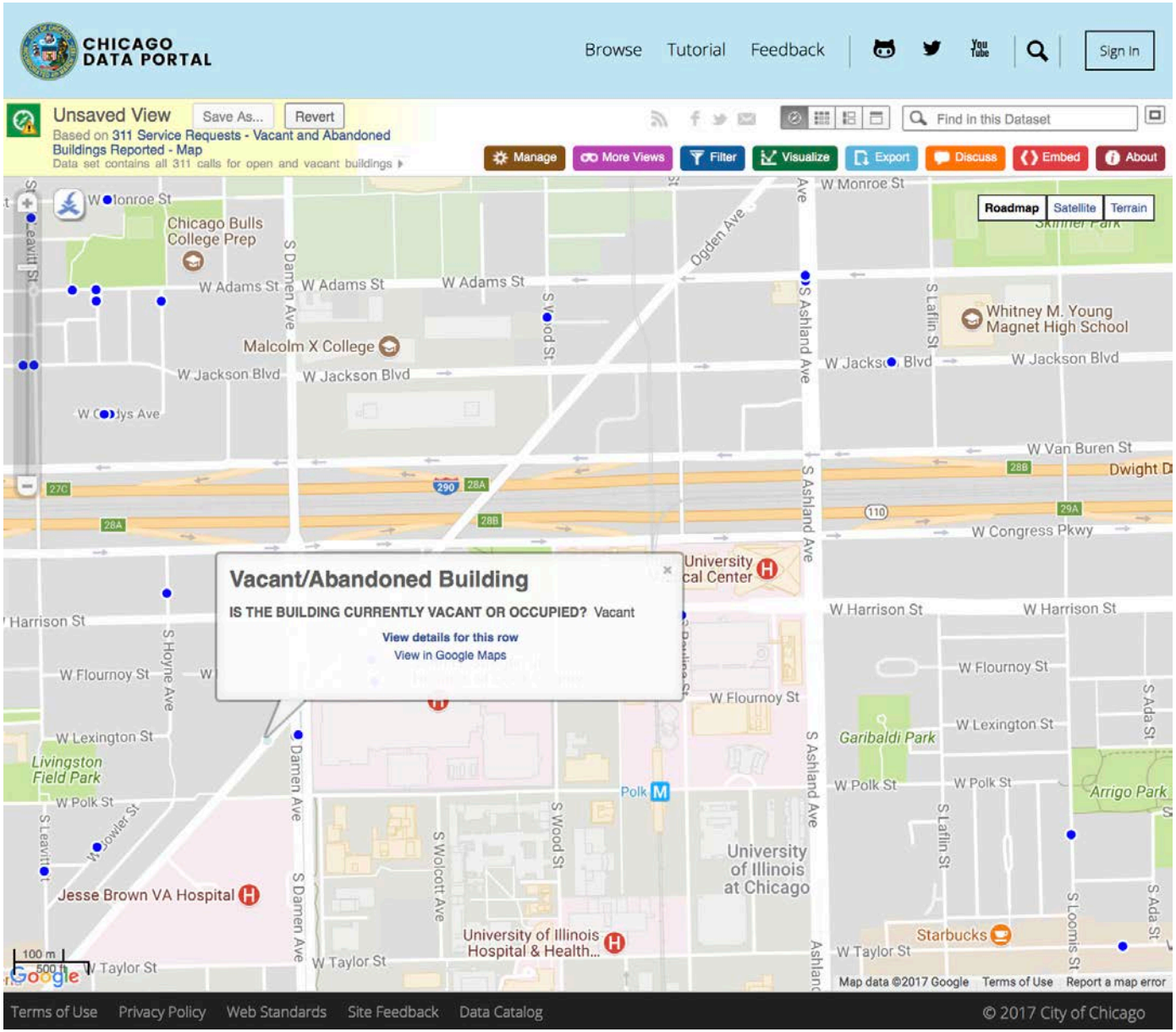
    prediction rather than explanation

- Big Data for the Social Sciences
  - new and big (or not so big) data sources
    - ubiquitous sensors - smart cities
    - open data portals - administrative data
    - social media data - Twitter analytics
    - 311 calls
    - cell phone data
  - geo-located and time stamped

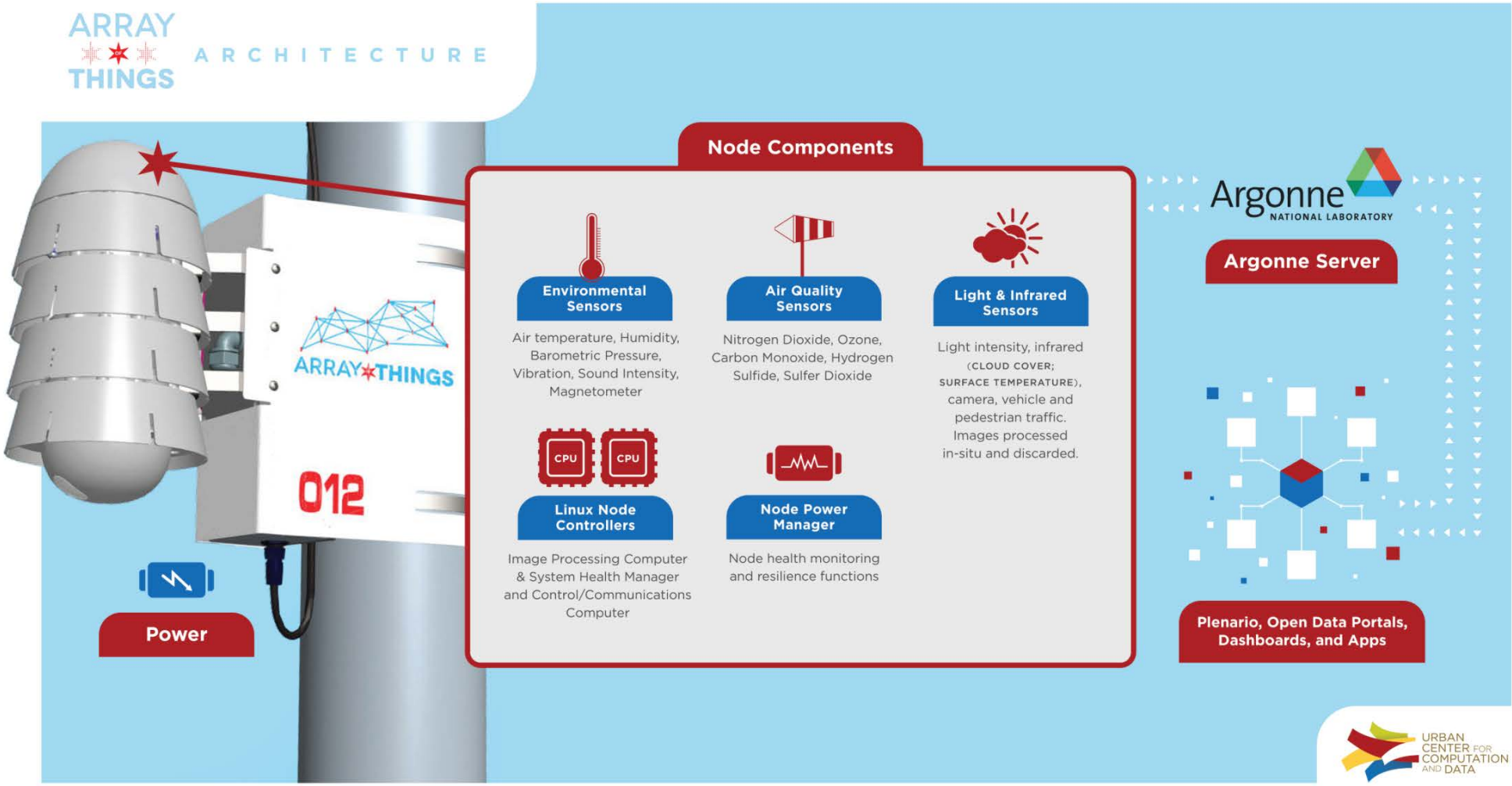
Some Examples



City of Chicago open data portal
https://data.cityofchicago.org

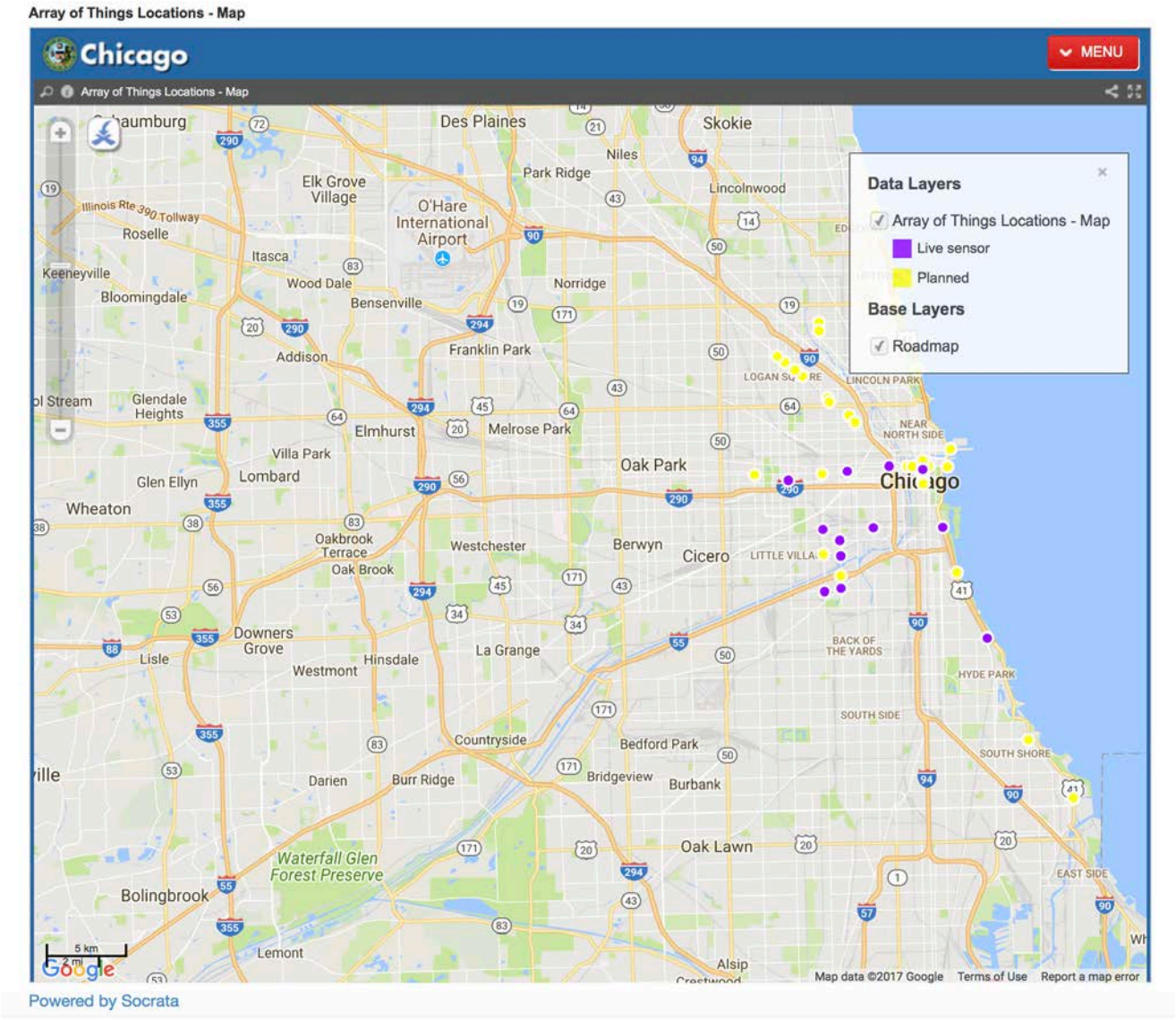


311 calls - abandoned buildings

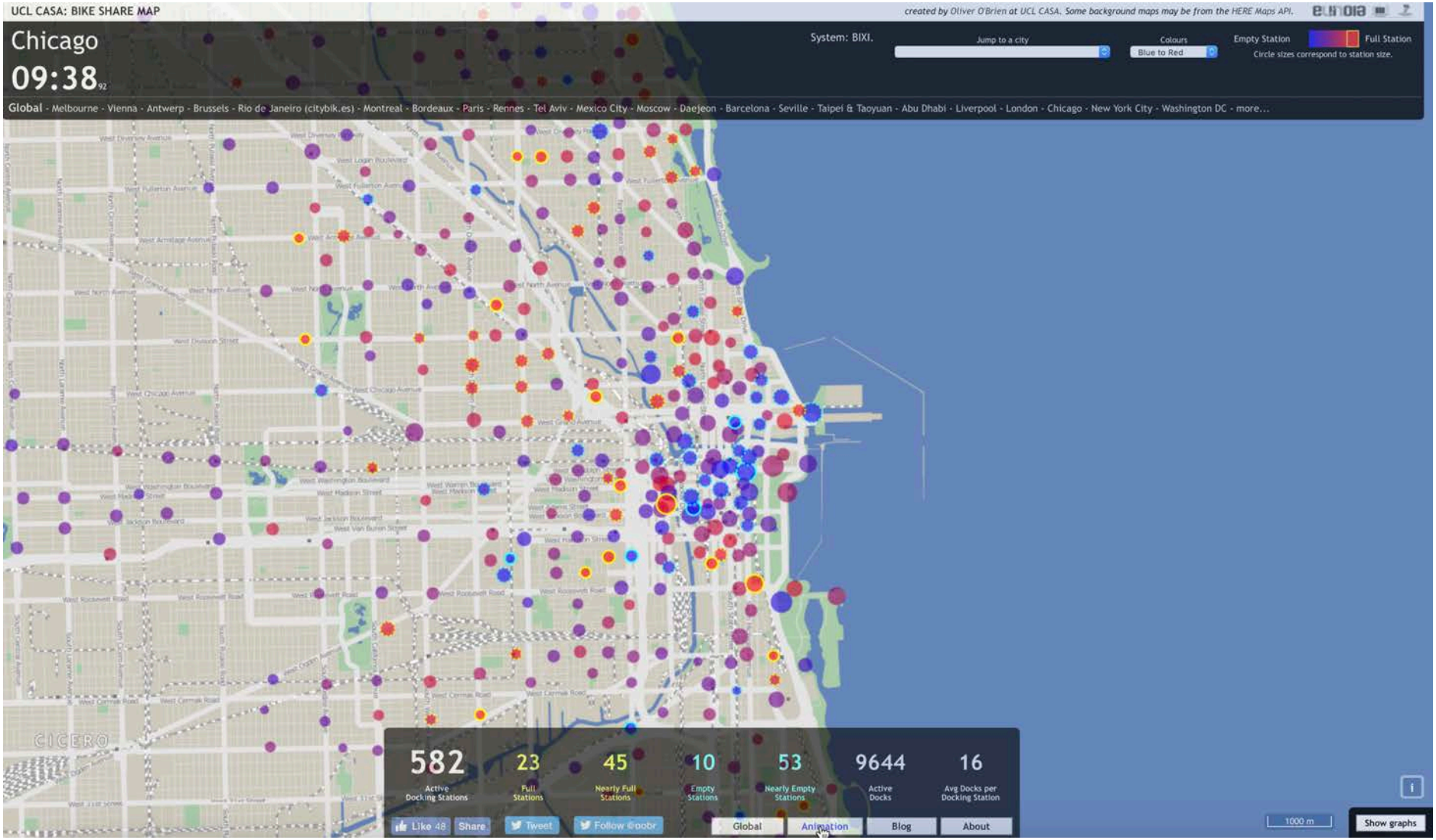


array of things sensor network

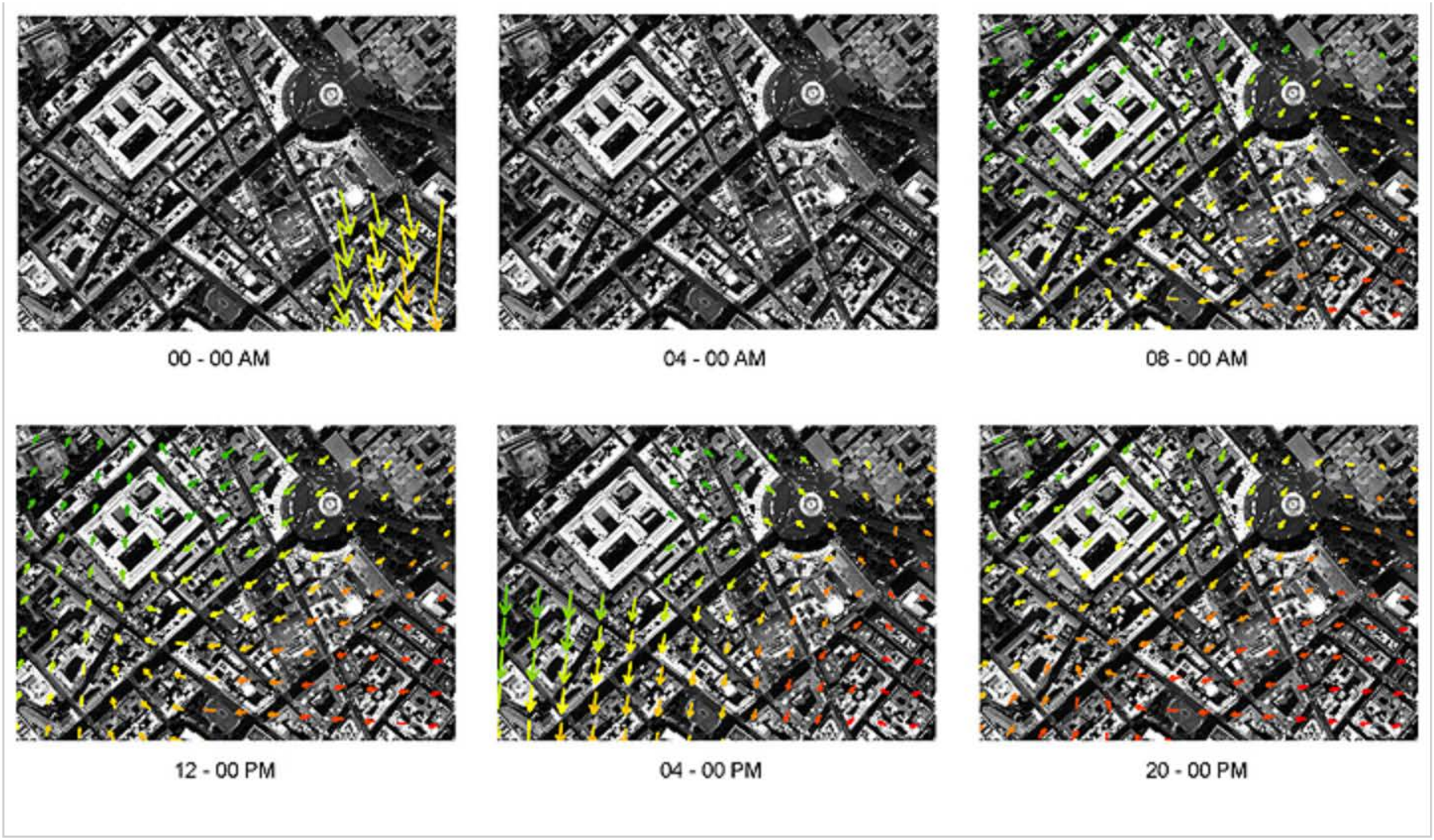
https://arrayofthings.github.io



array of things sensor locations

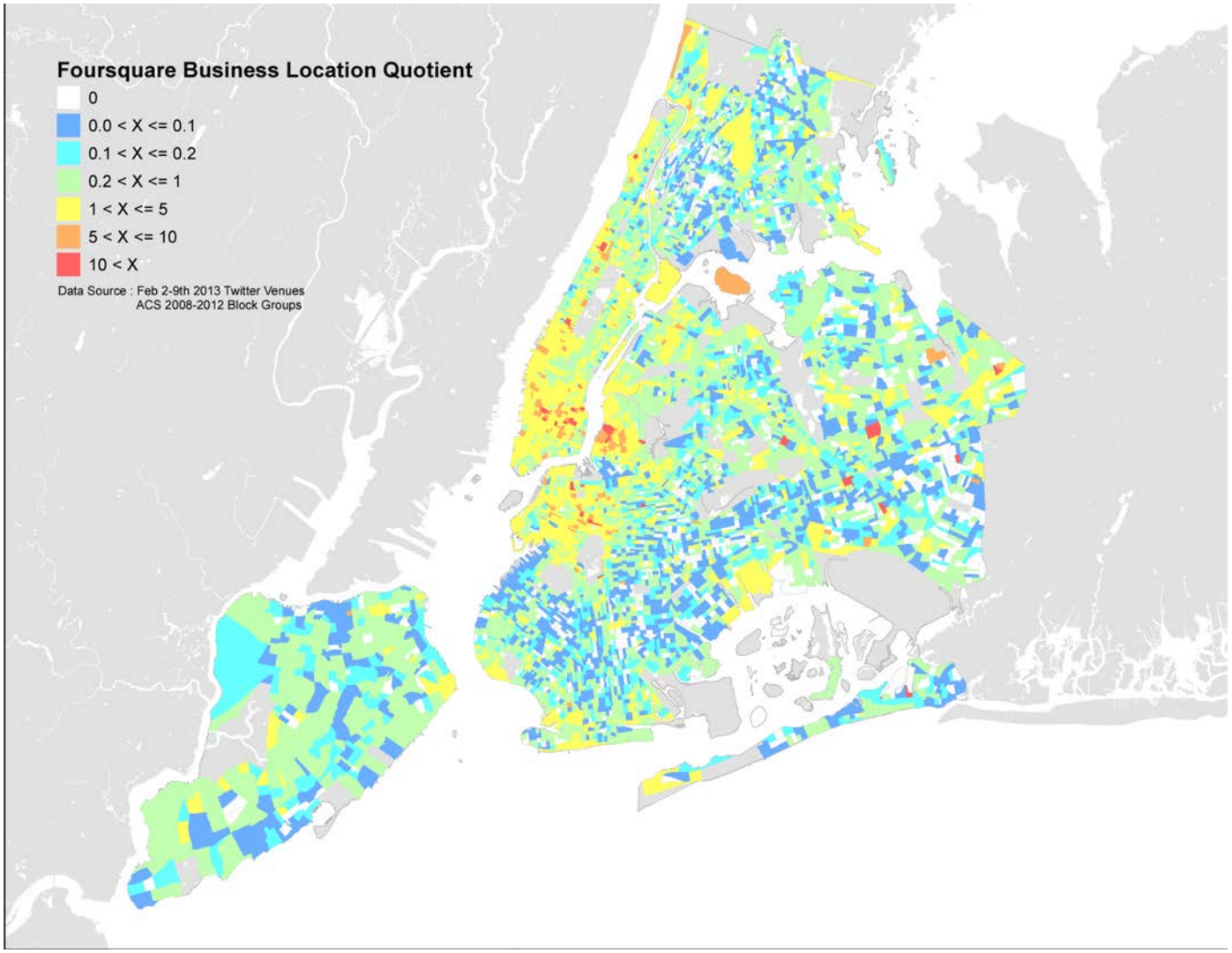


bike share locations



pulse of Rome - movement of cell phone calls

http://senseable.mit.edu/realtimerome



relative intensity of Foursquare check ins - NYC neighborhoods

Anselin and Williams (2016) Journal of Urbanism

Analytic Paradigm

- Computational Social Science

  computation as the third approach to scientific discovery

  new techniques

    simulation

    machine learning, data mining

    visual data exploration

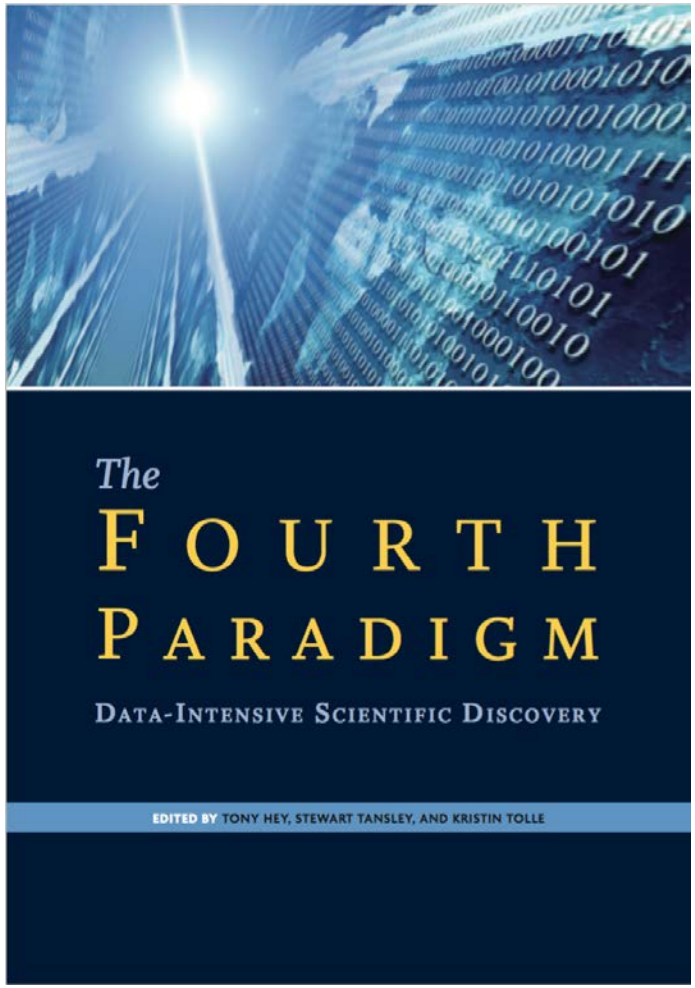
- Data Driven Science

  The Fourth Paradigm

  massive new data sets

  many social science applications driven by industry

    web marketing, recommender systems, etc.



SOCIAL SCIENCE

Computational Social Science

A field is emerging that leverages the capacity to collect and analyze data at a scale that may reveal patterns of individual and group behaviors.

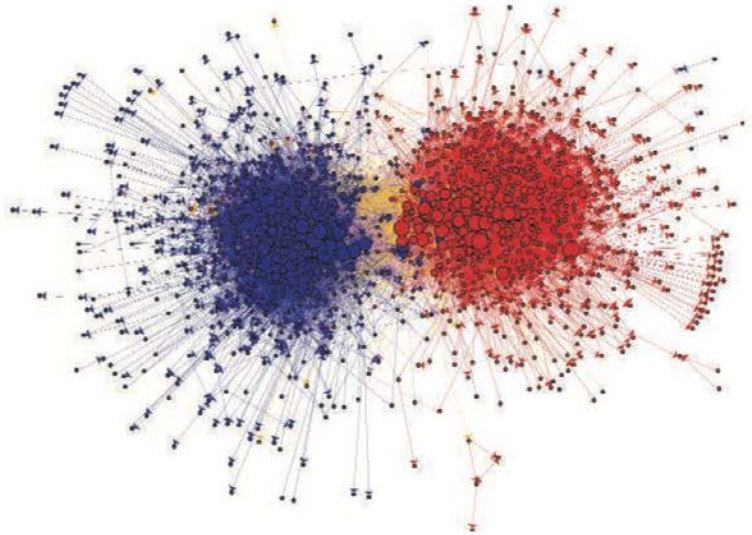
David Lazer,[1] Alex Pentland,[2] Lada Adamic,[3] Sinan Aral,[2,4] Albert-László Barabási,[5] Devon Brewer,[6] Nicholas Christakis,[1] Noshir Contractor,[7] James Fowler,[8] Myron Gutmann,[3] Tony Jebara,[9] Gary King,[1] Michael Macy,[10] Deb Roy,[2] Marshall Van Alstyne[2,11]

We live life in the network. We check our e-mails regularly, make mobile phone calls from almost any location, swipe transit cards to use public transportation, and make purchases with credit cards. Our movements in public places may be captured by video cameras, and our medical records stored as digital files. We may post blog entries accessible to anyone, or maintain friendships through online social networks. Each of these transactions leaves digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations, and societies.

The capacity to collect and analyze massive amounts of data has transformed such fields as biology and physics. But the emergence of a data-driven "computational social science" has been much slower. Leading journals in economics, sociology, and political science show little evidence of this field. But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

What value might a computational social science—based in an open academic environment—offer society, by enhancing understanding of individuals and collectives? What are the

[1]Harvard University, Cambridge, MA, USA. [2]Massachusetts Institute of Technology, Cambridge, MA, USA. [3]University of Michigan, Ann Arbor, MI, USA. [4]New York University, New York, NY, USA. [5]Northeastern University, Boston, MA, USA. [6]Interdisciplinary Scientific Research, Seattle, WA, USA. [7]Northwestern University, Evanston, IL, USA. [8]University of California–San Diego, La Jolla, CA, USA. [9]Columbia University, New York, NY, USA [10]Cornell University, Ithaca, NY, USA. [11]Boston University, Boston, MA, USA. E-mail: david_lazer@harvard.edu. Complete affiliations are listed in the supporting online material.

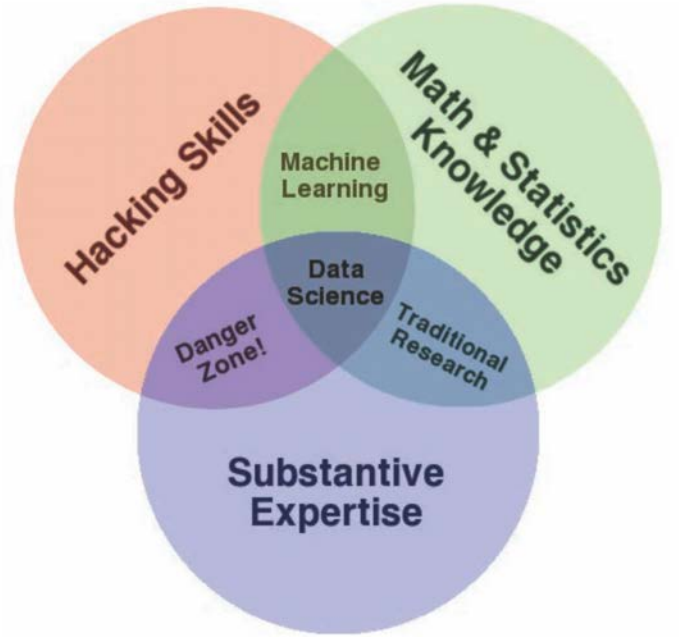


Data from the blogosphere. Shown is a link structure within a community of political blogs (from 2004), where red nodes indicate conservative blogs, and blue liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it. [Reproduced from (*8*) with permission from the Association for Computing Machinery]

- Data Science

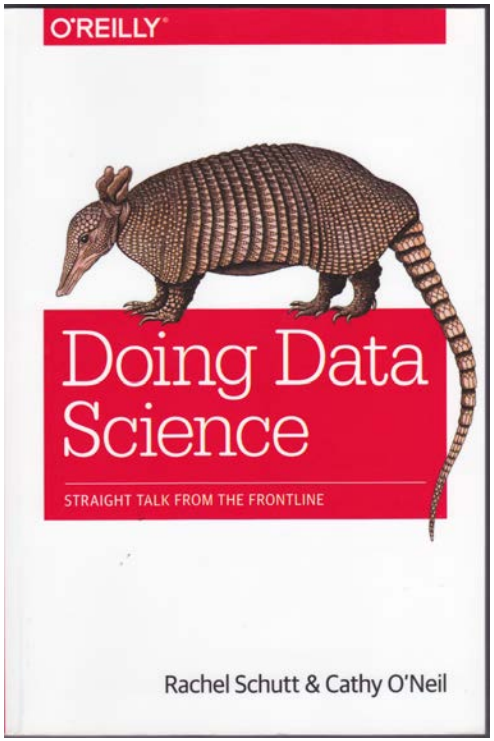
Nolan - Tempe Lang, Data Science in R (2015)

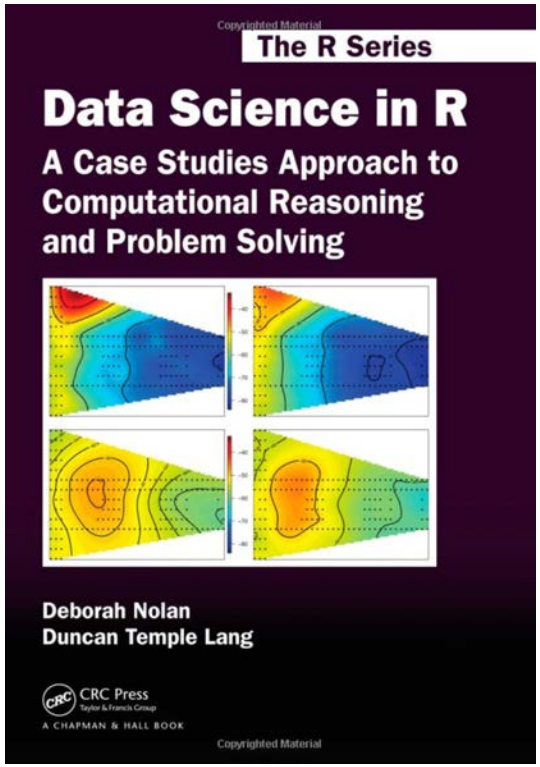
data science consists of "statistical computing and how to access, transform, manipulate, explore, visualize and reason about data"

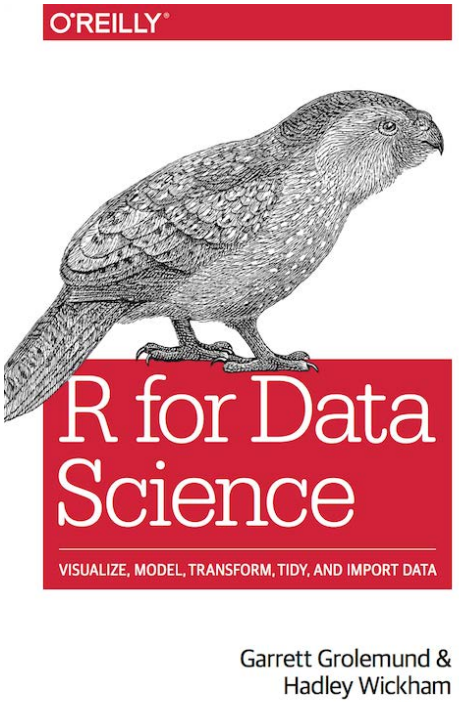


Source: Drew Conway

data science Venn diagram







selected operational data science texts

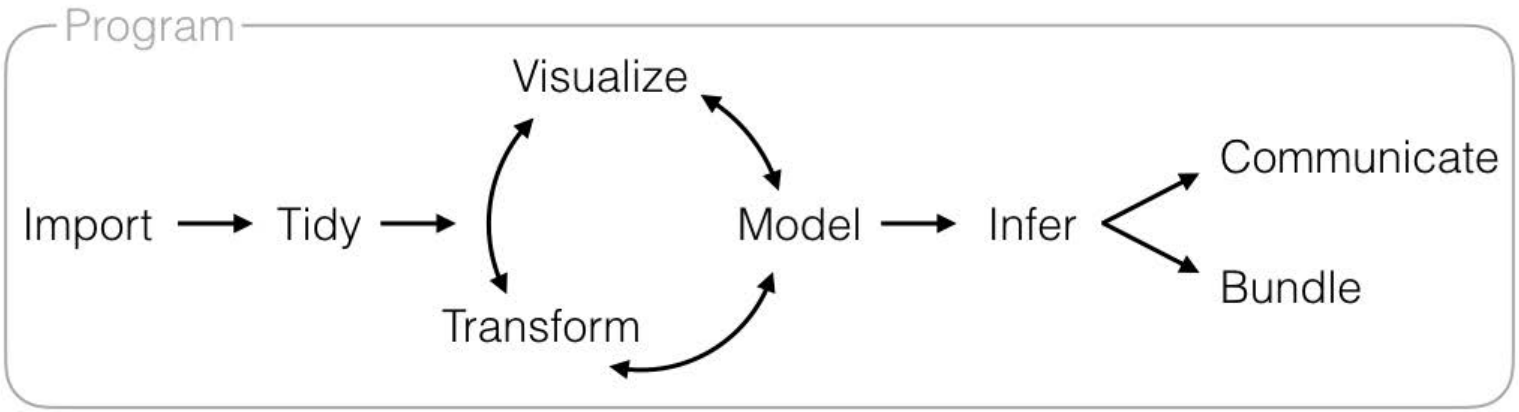
- Spatial Data Science

  - explicit treatment of spatial aspects

    integration of geocomputation, spatial statistics, spatial econometrics, exploratory spatial data analysis, visual spatial analytics, spatial data mining, spatial optimization

    80% effort is data preparation (Dasu and Johnson 2003)

      algorithms, data structures, workflow



data science process

Grolemund and Wickham (2017)

- What's Involved in Spatial Data Science?

  data manipulation (munging, wrangling)

  data integration

  data exploration, pattern recognition, associations

  visualization

  modeling (prediction and explanation), classification, simulation, optimization

- lots of different software tools

Example

- Digital Neighborhoods (with Sarah Williams)

  twitter and foursquare locations in NYC

  first week of Feb 2014

  573,278 tweets and 589,091 foursquare check-ins

- Data Manipulation

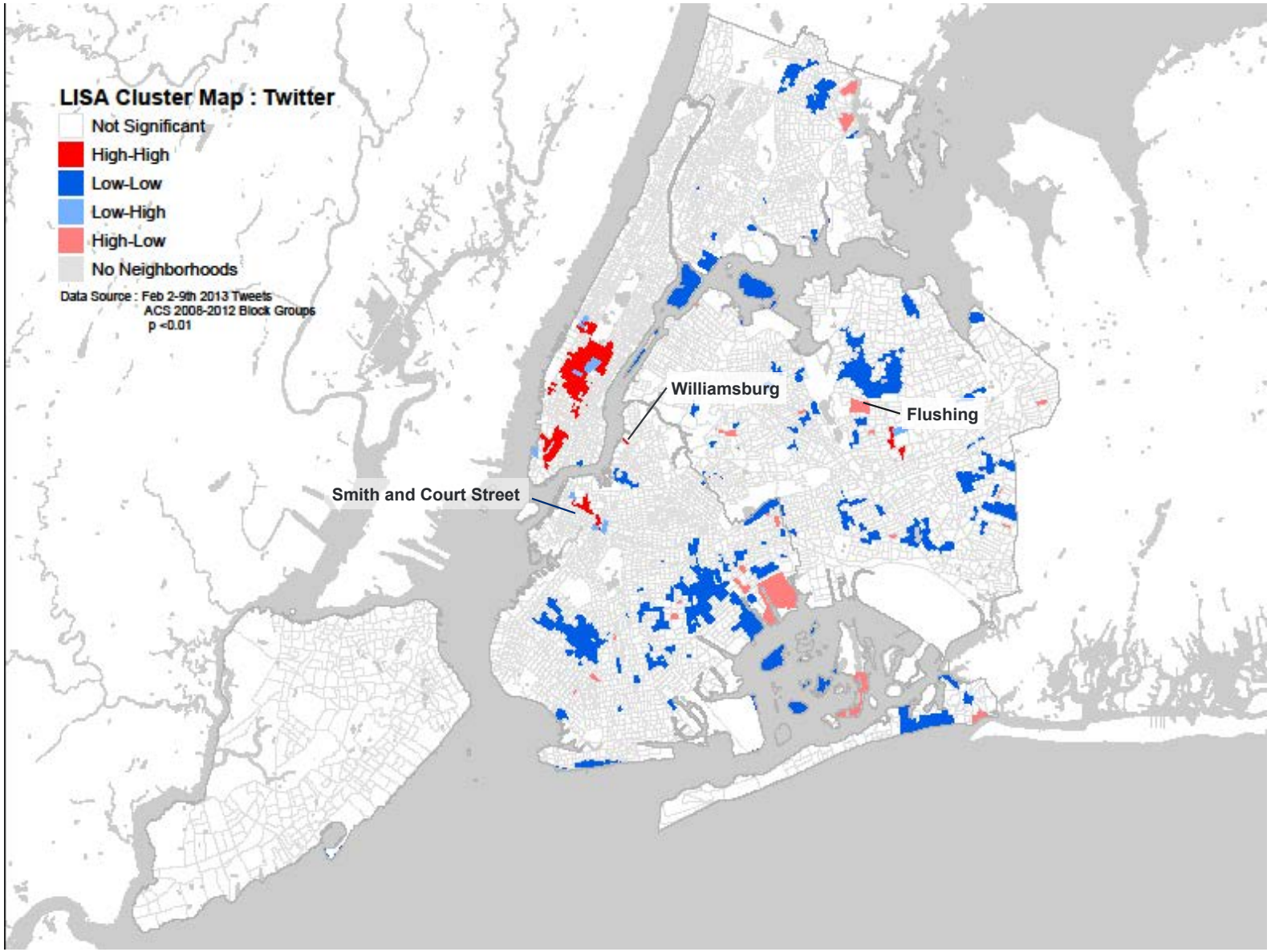
  parse Twitter JSON files and convert to csv

    5760 files, more than 5 million messages
    > 20Gb of memory (Python or R)

  convert text file to spatial data base (PostGIS)

  spatially aggregate points to block group totals
  (n = 6454) (PostGIS or R)

  run local Moran statistics + visualize (GeoDa)



Spatial Data Types and Research Questions

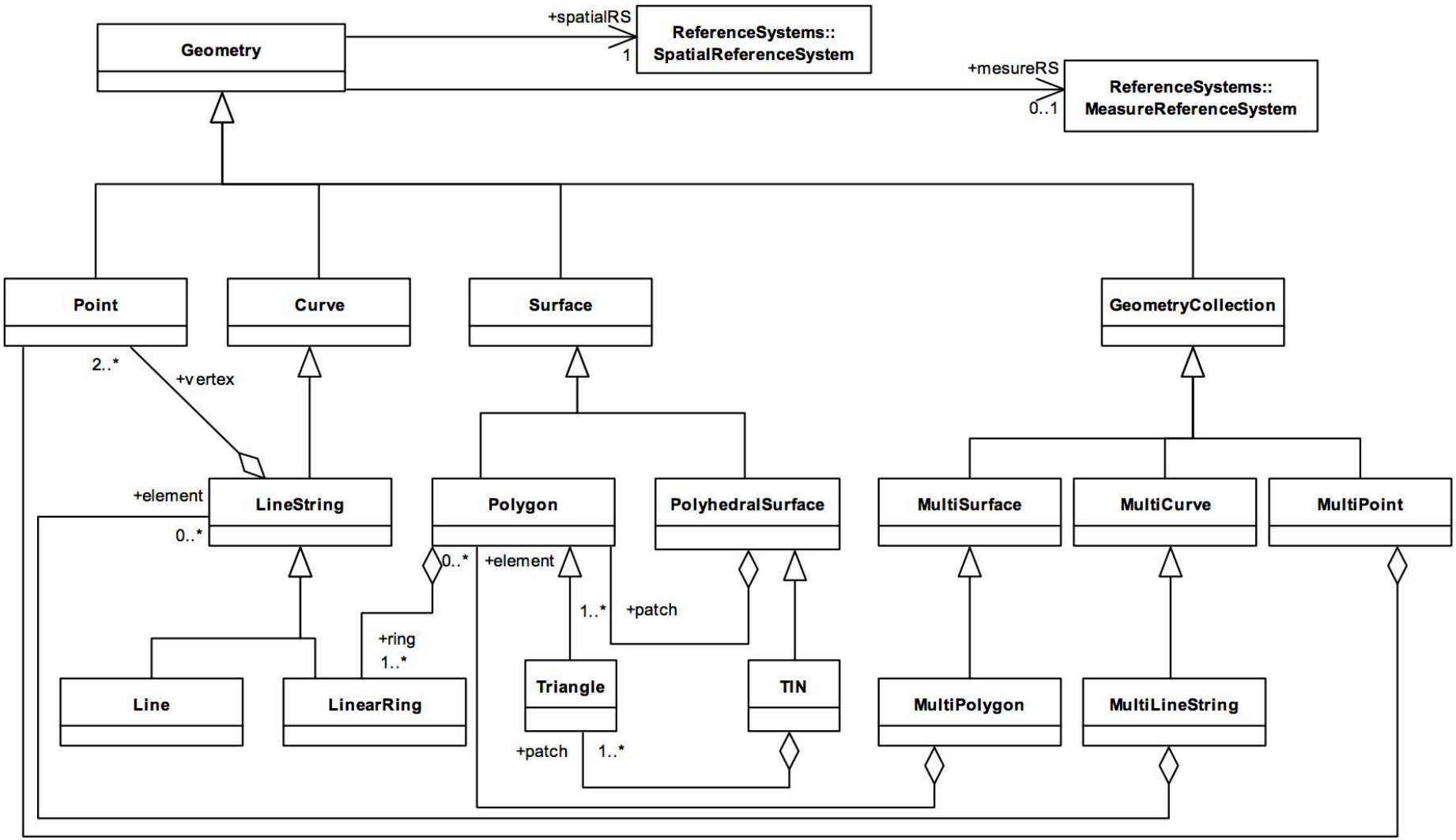
- Spatial Data Structures

  formal representation of geographic features

  abstracted to points, lines and polygons

  spatial databases

    spatial index: speed up search



OGC Simple Features Specification

- Spatial Data Types for Analysis
  - points
    - location as a random event
    - locations of crimes, accidents, grocery stores
  - surfaces
    - continuous spatial field
    - air quality surface, noise surface, price surface

- Spatial Data Types for Analysis (2)

  discrete spatial data - lattice data

    areal units

    census tracts, counties, countries

  networks

    nodes and links

    street network, river network, social network

- Space-Time Data

  fixed spatial locations over time

  - time-in-space

    panel data = pooled cross-section and time series

    e.g., crime by neighborhood over time

- Space-Time Data (2)

  changing spatial locations over time

  - space-in-time

    moving objects

    e.g., taxi with GPS, cell phone calls, animal tracking

- Data Types and Data Analysis

  the type of data determines what analysis can be carried out

  types of research questions

  are traffic accidents located randomly in space or clustered > point pattern analysis

  given sensor measurements on air quality, what is an air quality surface for a region > spatial interpolation

  where are hot spots of mortgage foreclosure in the city > cluster detection

  how are house prices affected by unobserved neighborhood effects > spatial regression

- Some Important Characteristics

  are the data sampled (e.g., sensor locations) or is it the population (e.g., all the census tracts)

  are the spatial units discrete (areal units) or continuous (surface)

  are the locations given (e.g., areal units) or themselves random (e.g., location of events)

Pitfalls

- Ecological Fallacy

  individual behavior cannot be explained at the aggregate level

  issue of interpretation

    e.g., county homicide rates do not explain individual criminal behavior

  model aggregate dependent variables with aggregate explanatory variables

  alternative: multilevel modeling

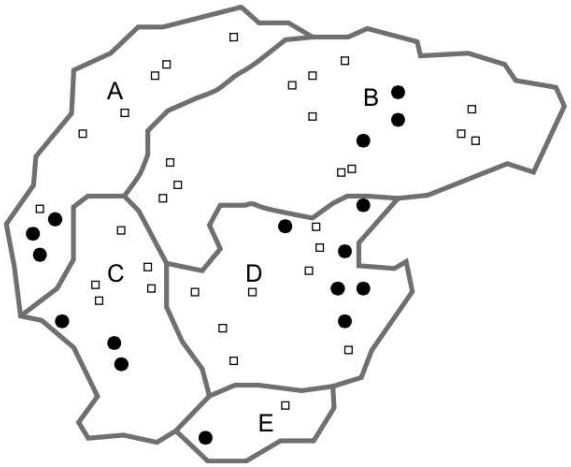
- Modifiable Areal Unit Problem (MAUP)

  what is the proper spatial scale of analysis?

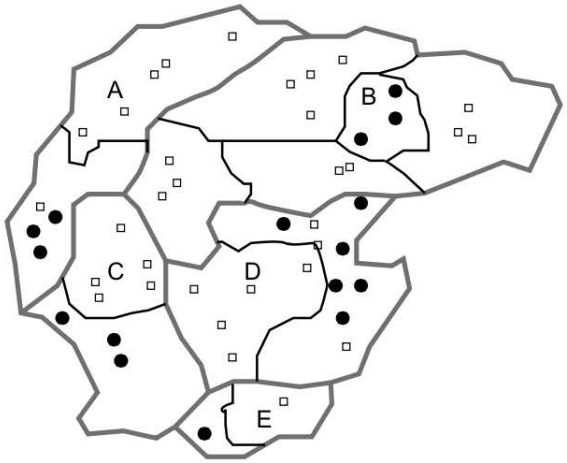
  a million spatial autocorrelation coefficients (Openshaw)

  spatial heterogeneity - different processes at different locations/scales

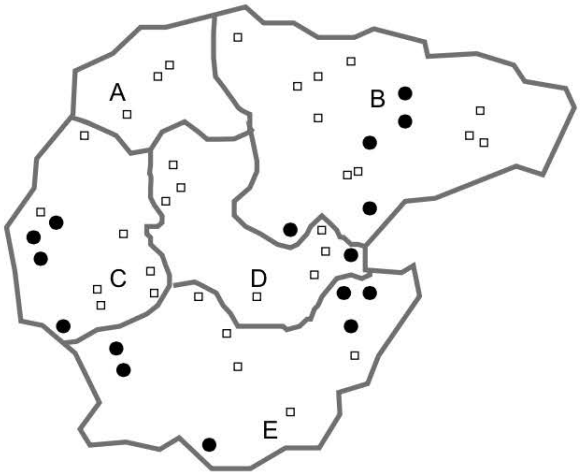
  both size and spatial arrangement of spatial units matter



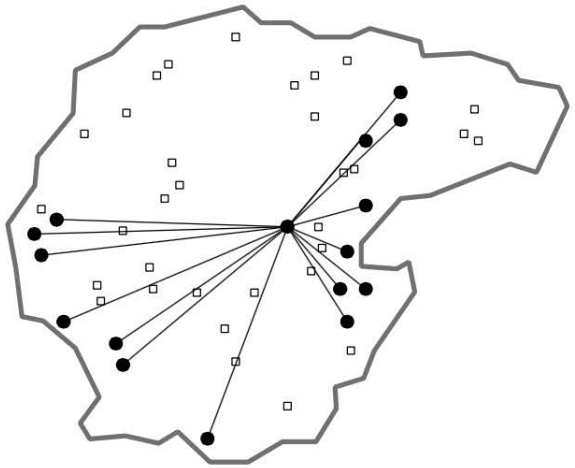
(a) Hypothetical country with 5 regions



(b) Scaling problem: Change of aggregation level



(c) Zoning problem: Change of boundaries



(d) MAUP-free distance based approach

Source: Scholl and Brenner (2012)

- Change of Support Problem (COSP)

  variables measured at different spatial scales

    nested, hierarchical structures

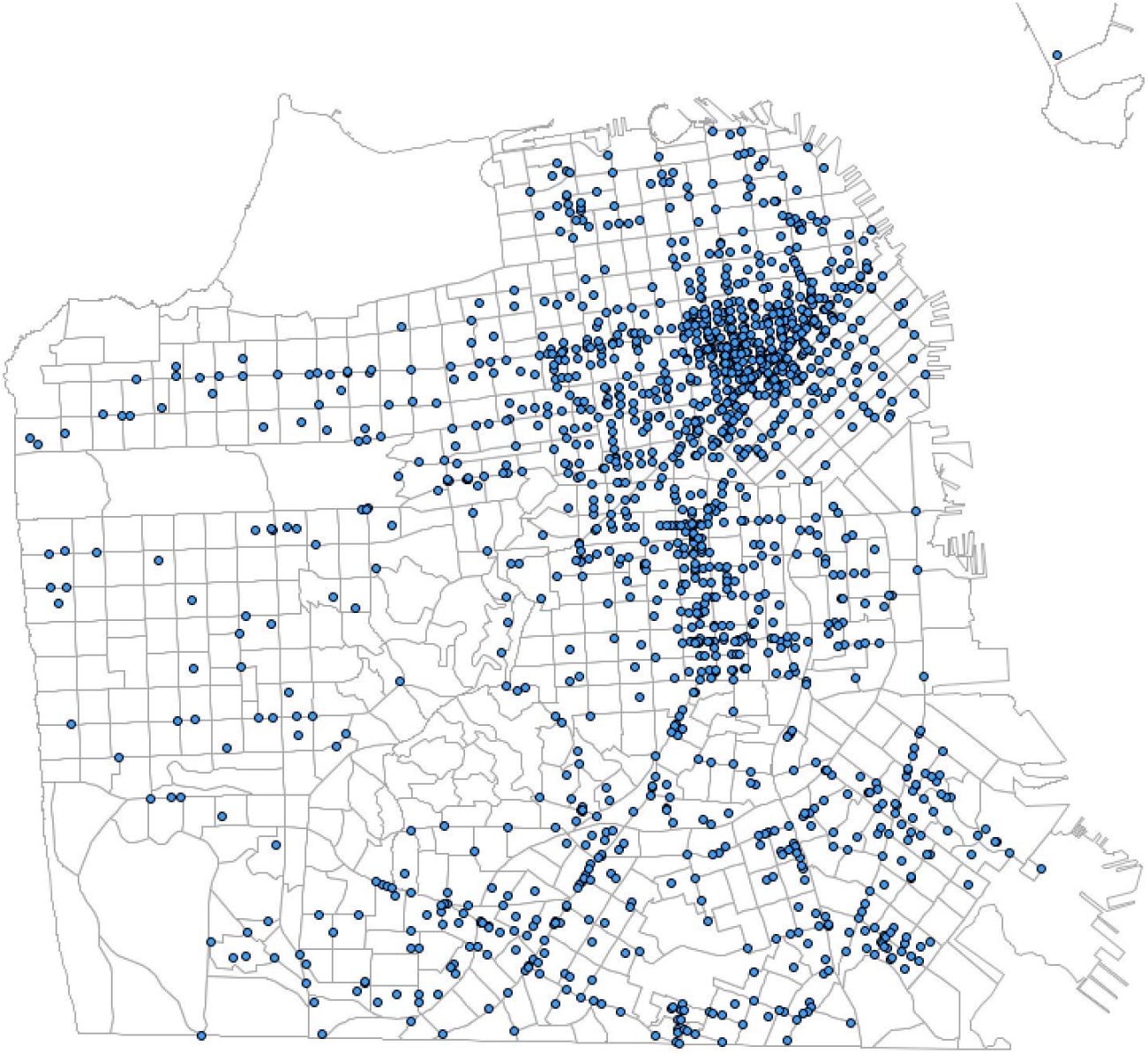
    non-nested, overlapping

  solutions

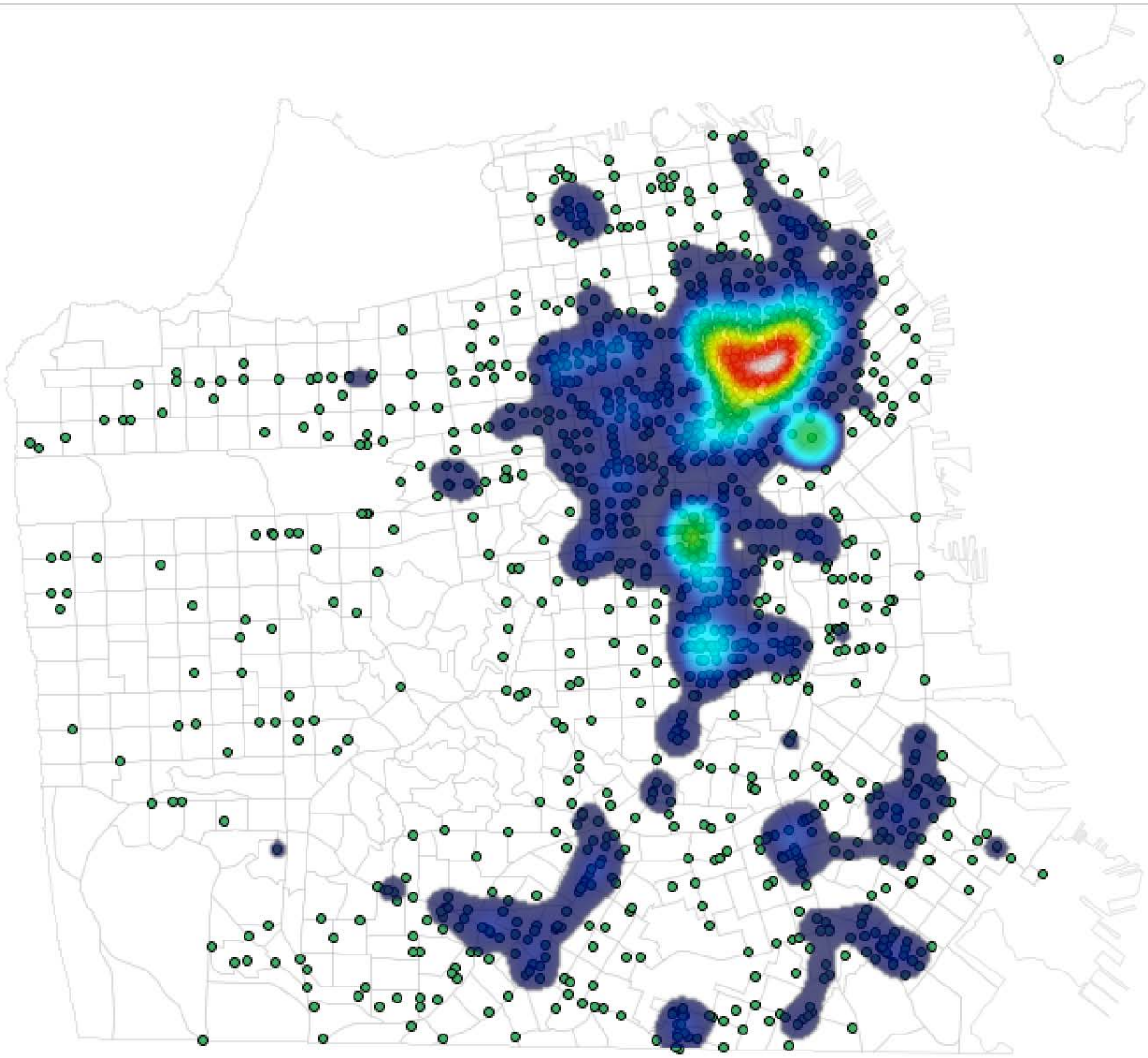
    aggregate up to a common scale

    interpolate/impute - Bayesian approach

Examples

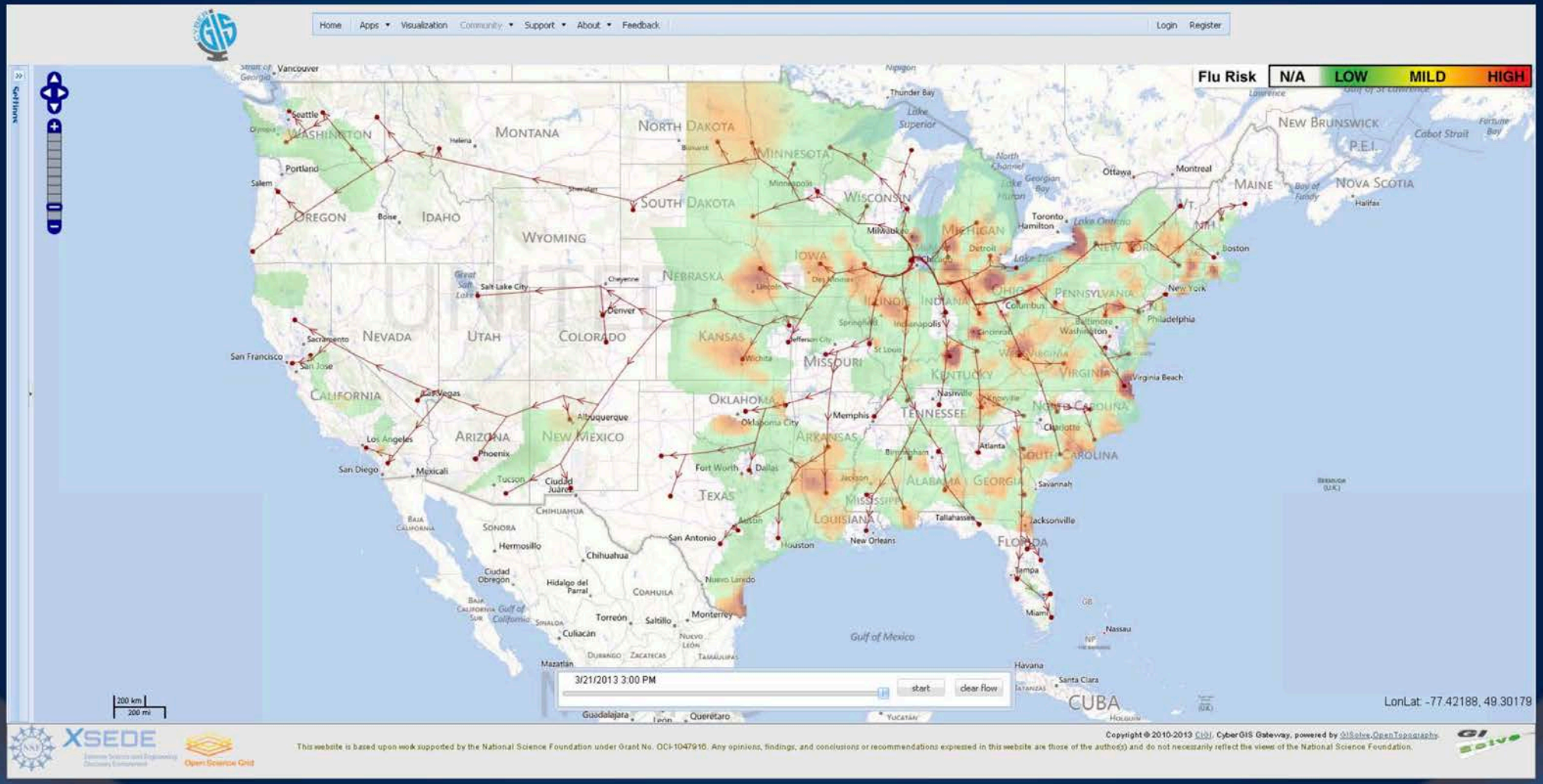


Events: Car thefts in San Francisco

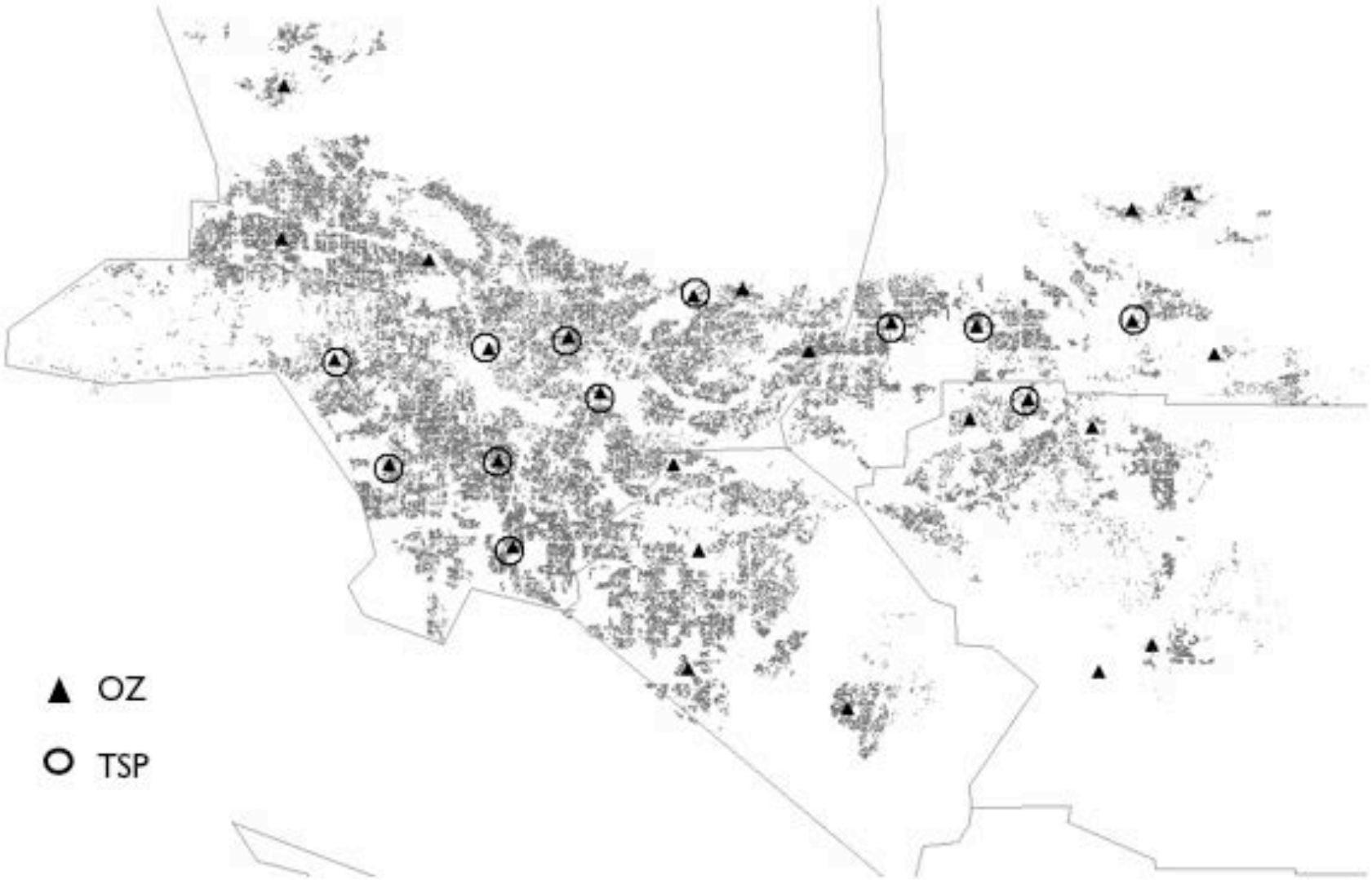


cluster detection
San Francisco car thefts - Heat Map (KDE)

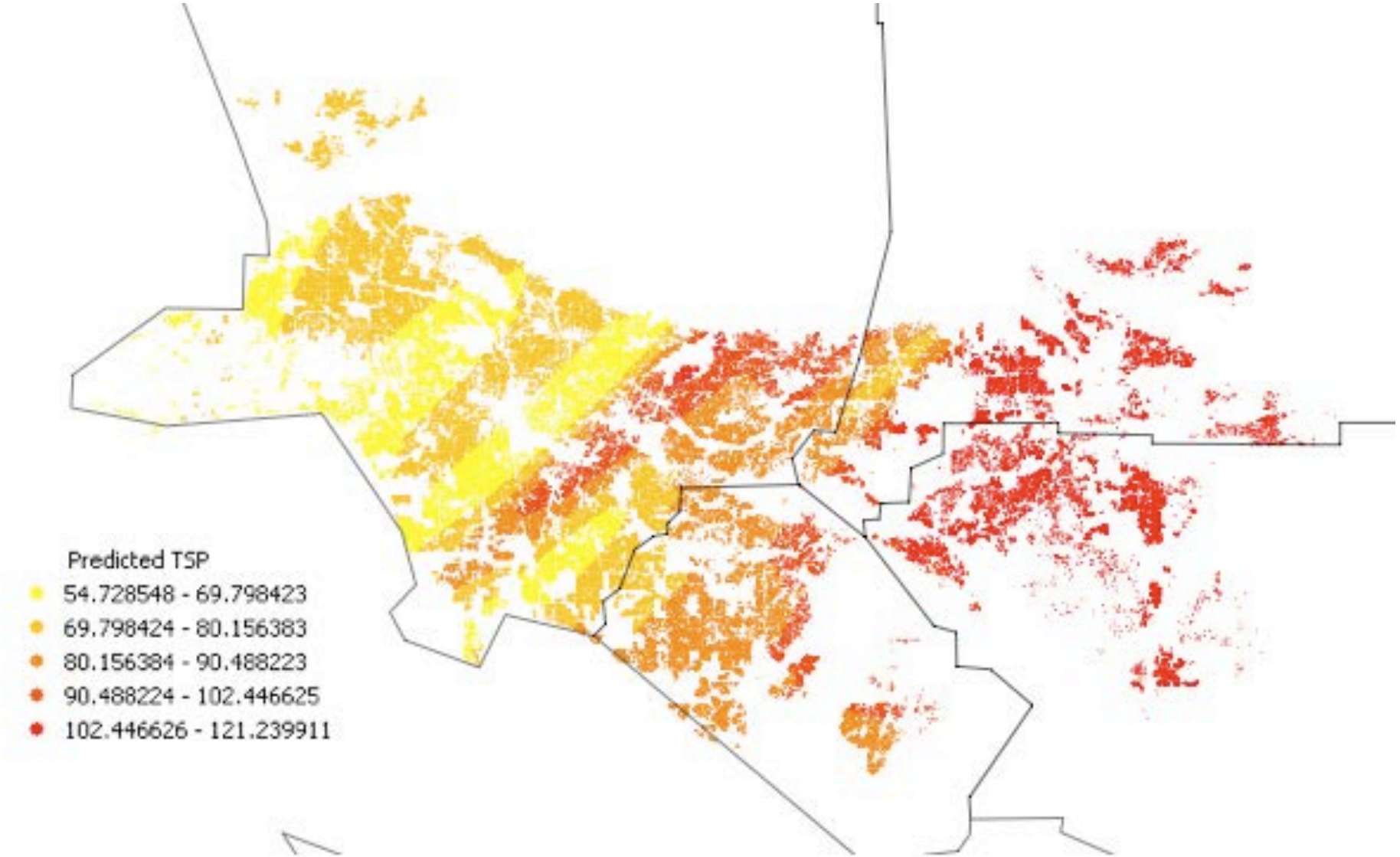
http://flumapper.org/



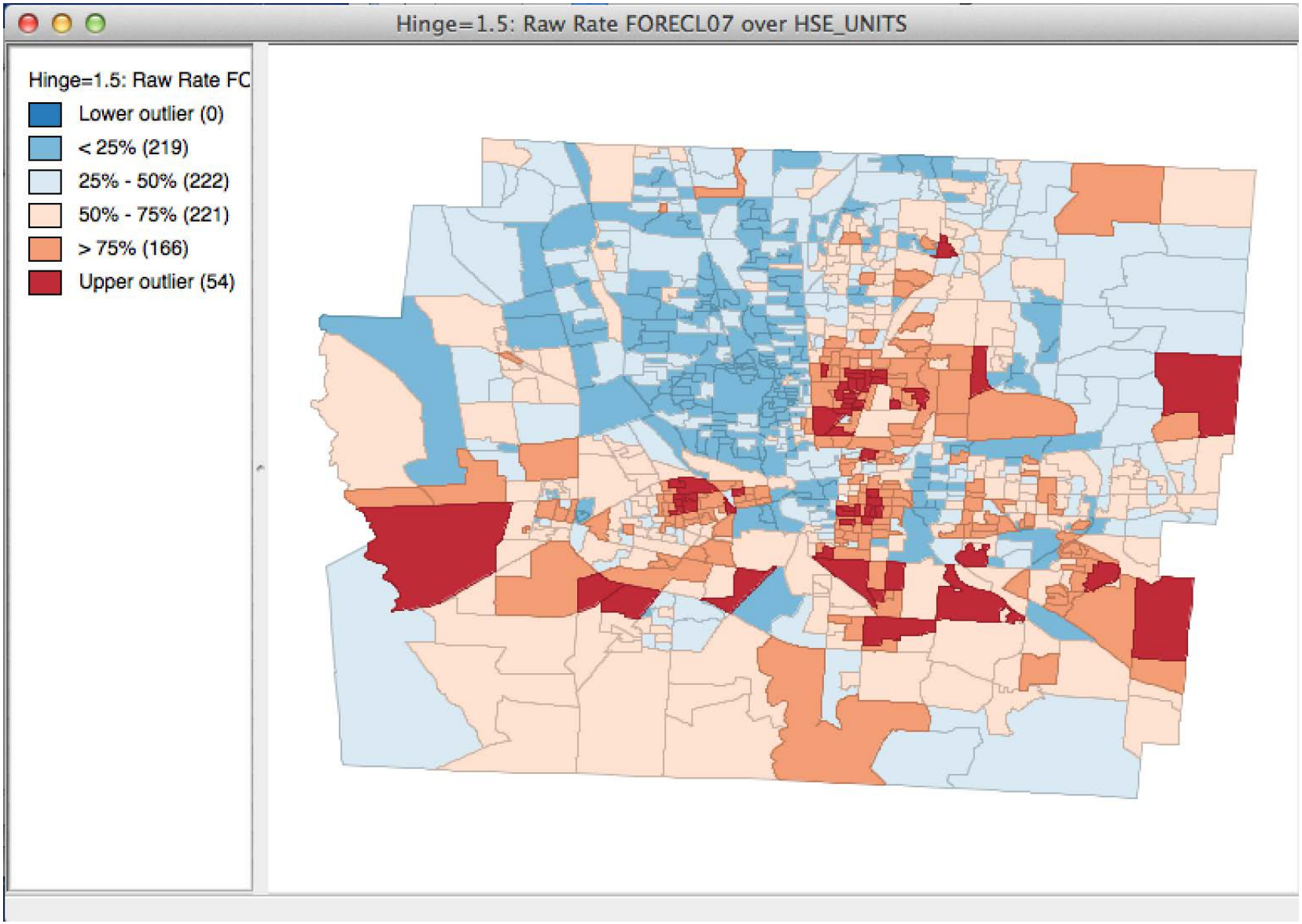
flu hot spots and flows from tweets



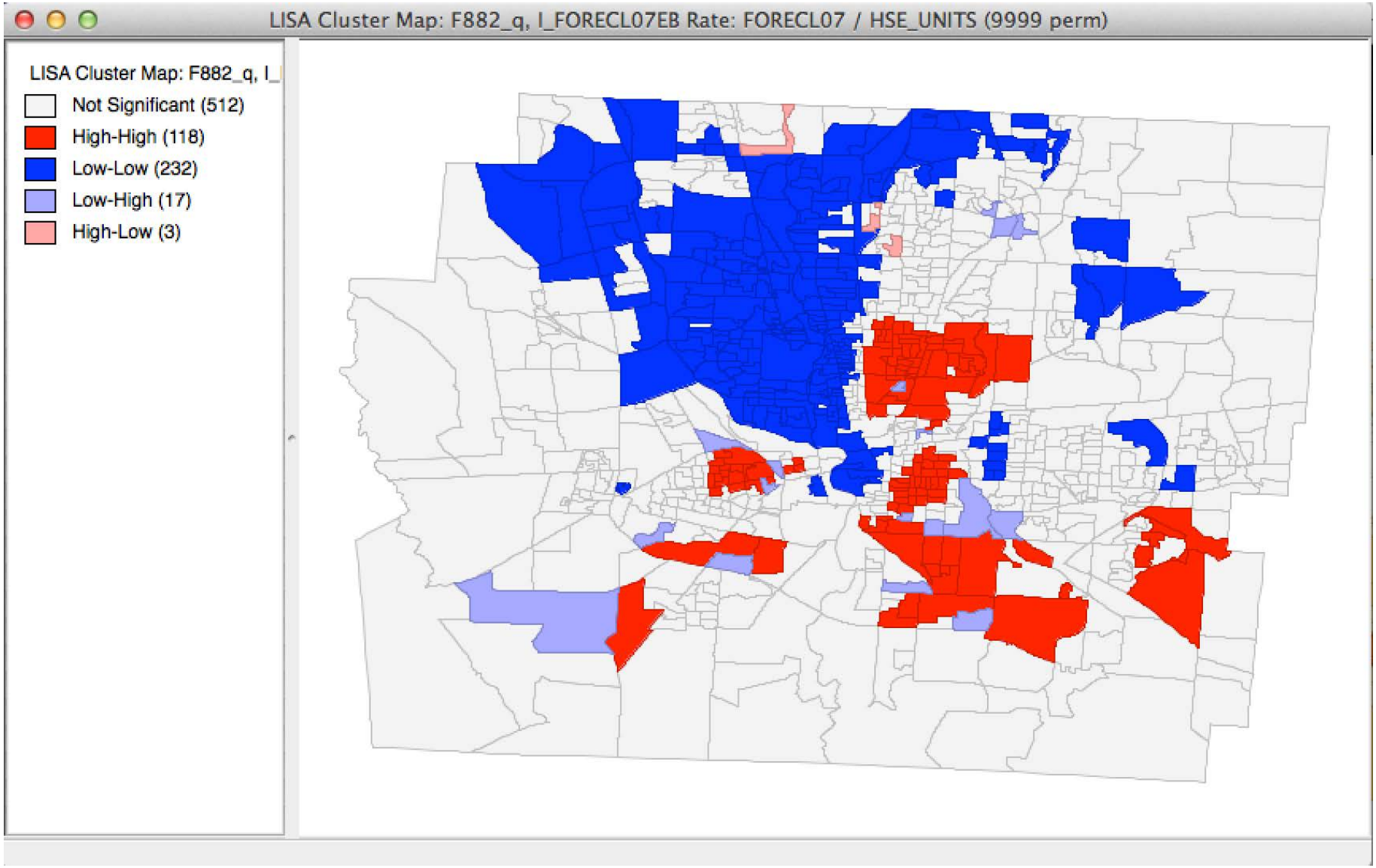
LA Basin air quality monitoring stations



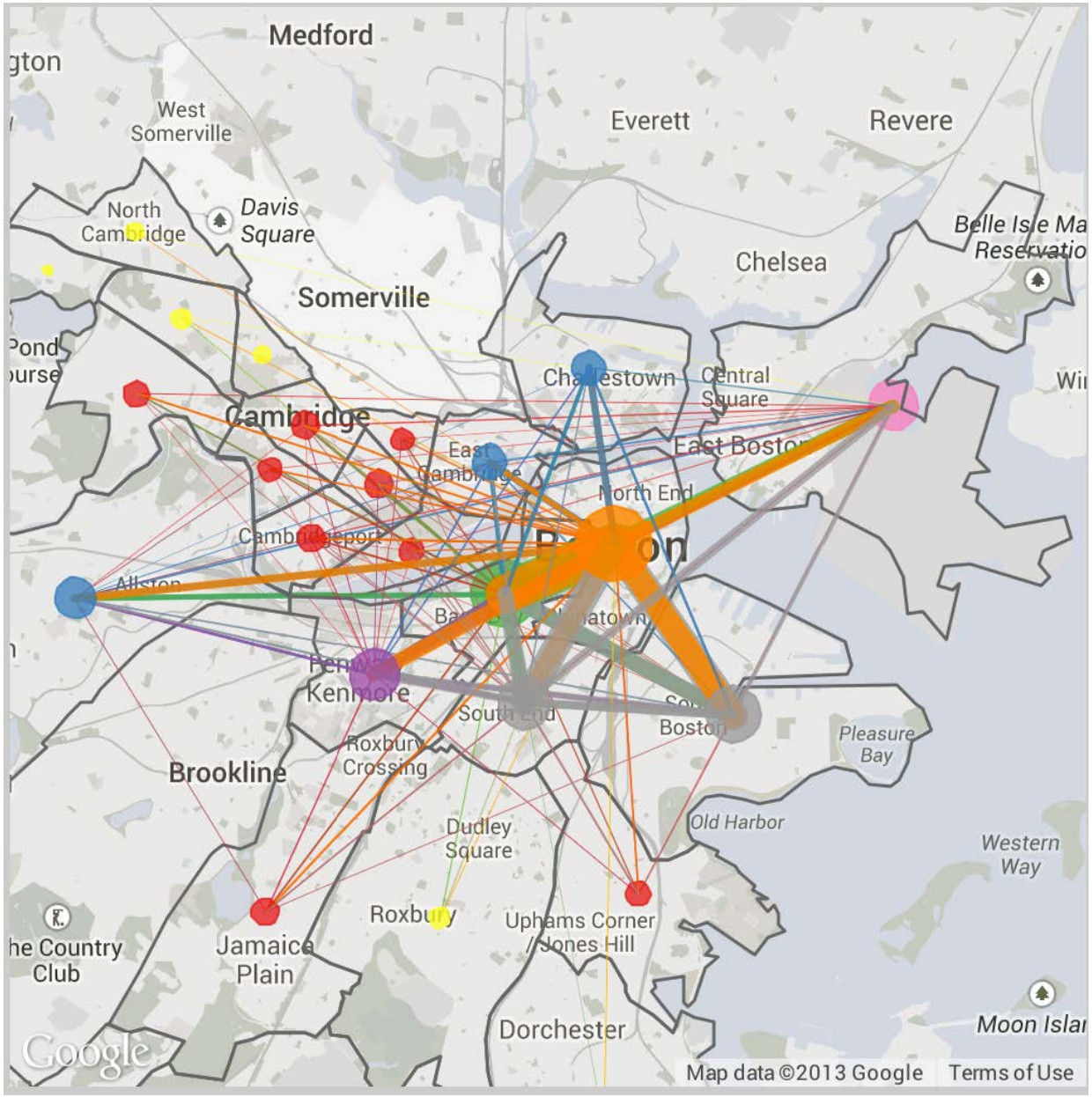
Interpolated air quality surface (Kriging) - pm10



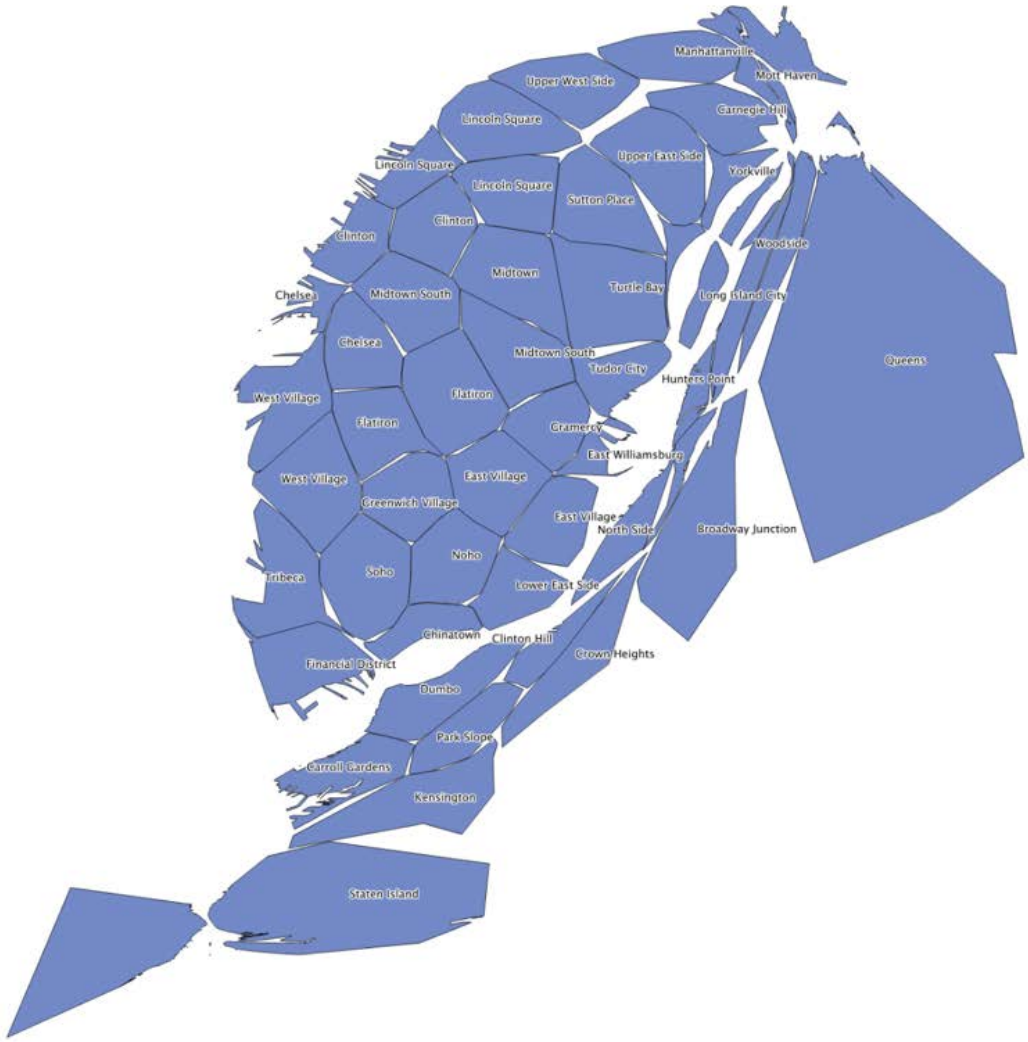
Franklin County census tract foreclosure rate outlier map



Local Moran cluster map
Franklin county census tract foreclosure rate



uber Boston ridership flows



uber ridership cartogram (NYC)