Transforming Open Data into Insights

Darren Erik Vengroff
Rachael Weiss Riley || Erin Stein || Kaushik Mohan

How Open Data Can Help Us Reimagine NYC Neighborhoods

#nyc || #opendata || #NewerHoods

@tsdataclinic
Disclaimer

This document is being distributed for informational and educational purposes only and is not an offer to sell or the solicitation of an offer to buy any securities or other instruments. The information contained herein is not intended to provide, and should not be relied upon for, investment advice. The views expressed herein are not necessarily the views of Two Sigma Investments, LP or any of its affiliates (collectively, “Two Sigma”). Such views reflect the assumptions of the author(s) of the document and are subject to change without notice. The document may employ data derived from third-party sources. No representation is made by Two Sigma as to the accuracy of such information and the use of such information in no way implies an endorsement of the source of such information or its validity.

The copyrights and/or trademarks in some of the images, logos or other material used herein may be owned by entities other than Two Sigma. If so, such copyrights and/or trademarks are most likely owned by the entity that created the material and are used purely for identification and comment as fair use under international copyright and/or trademark laws. Use of such image, copyright or trademark does not imply any association with such organization (or endorsement of such organization) by Two Sigma, nor vice versa.
Who We Are

An intro to Two Sigma and Data Clinic
- Pro bono data science and engineering support
- Partner with nonprofits, government agencies, and academic institutions
- Volunteer teams staffed by Two Sigma employees
- Self-guided research to contribute to the data-for-good movement

DATA CLINIC

Est. 2014
Corporate Data Philanthropy (CDP)

Leveraging a company’s people, data, and technology for social benefit
Partner Projects

1. Engagement + Team
2. Research + Development
3. Results + Impact
Example Partner Projects

Can we detect water leaks & meter malfunctions based on a customer’s previous usage?

What is causing the U.S. urban-rural jail incarceration divide?

How can we better guide participants toward programs that will increase their likelihood of success?
Common Threads

✓ Established organizations

✓ A lot of data in-house

✓ Research questions that could be answered by in-house data
Open Data

An under-utilized resource
Why Use Open Data?

→ It exists when in-house data doesn't!
→ Open data is diverse
→ Varied applications/use cases
  ◆ Add value to daily operations
  ◆ Advance research
  ◆ Build business case for data strategy
Building a Proof of Concept

→ What predicts future oil and gas industry violations?

→ Past violations + inspection frequency were highly predictive of future violations

→ Resulted in:
  ♦ Culture shift at EDF
  ♦ Shared, inter-organizational research strategy
Challenges of Open Data

- Being publicly available != being “open”
- Having a “unique” ID might not be enough
- Limitations in the availability of open data may hinder timely analyses
- Understanding the original purpose of different data sets
- Efforts and solutions are often in isolation
Advocate for and contribute to open data
...for all!
NewerHoods

Redefining NYC neighborhood boundaries using open data
NYC ♥ Neighborhoods

THE LEAST IDEAL NEIGHBORHOOD FOR MAKING OUT
HIGH BRIDGE, BRONX

BASSED ON THE NUMBER OF GARLIC-BASED ORDERS.

seamless
HOW NEW YORK EATS

THE NEIGHBORHOOD WITH THE IDENTITIEST
UPPER EAST SIDE

BASED ON CHICAGO-STYLE STEAKS

High-quality, Award-winning.

Find your place.

StreetEasy
Search NYC Apartments

PLACES WHERE I CAN STILL SAY, "I LIVE IN TRIBeca."

TRIBeca

DATA CLINIC
Neighborhoods are Newsworthy

See the NYC neighborhoods where displacement is a growing threat

This interactive map illustrates the various factors that contribute to displacement across various neighborhoods

By Ameena Walker | Oct 2, 2018, 4:18pm EDT
### What is a Neighborhood?

<table>
<thead>
<tr>
<th>Social common perceptions</th>
<th>Administrative who is served</th>
<th>Statistical data collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>- No clear definition</td>
<td>- Defined specifically to serve the respective organization</td>
<td>- Defined to capture areas with a specific population count for data collection &amp; organization</td>
</tr>
<tr>
<td>- Varies from person to person</td>
<td>- Optimized based on organizational costs</td>
<td>- Updated approximately every 10 years</td>
</tr>
<tr>
<td>- Eg: “hip” neighborhoods in Brooklyn</td>
<td>- Static definitions</td>
<td>- Eg: US Census block groups or tracts</td>
</tr>
</tbody>
</table>
How can we contribute??
Some Current Tools

(Citizens Housing and Planning Council)

(Environmental Systems Research Institute (ESRI))
An open, flexible, and dynamic tool
The Vision for NewerHoods

➔ Make publicly available data more accessible & insightful

➔ Driven by needs & interests of the community

➔ A tool for the community, sustained by the community
  ◆ Additional data sets
  ◆ Use cases
  ◆ Improved methodology
The Approach

**Open Data**
NYC Open Data to gather information on a variety of dimensions contributing to quality of life

**Local Attributes**
Extract multiple different attributes for every census tract from these data sets

**Clustering**
Use Machine Learning techniques to find homogenous areas based on chosen characteristics
What is a Census Tract?

Source: NYC Department of City Planning
Challenges

➔ Real-estate sales data is pretty hard to work with
  ◆ Requires merging 4 different data sets
  ◆ Mismatches in unique identifiers for certain types of buildings
  ◆ Extensive cleaning process

➔ Inconsistencies in spatial reference systems
  ◆ Need to spatially join different data sets
NewerHoods
FROM TWO SIGMA DATA CLINIC

Choose characteristics to draw neighborhoods.

HOUSING
☐ Age of buildings
☐ Median Sale Price

CRIME
☐ Violations
☐ Felonies
☐ Misdemeanors

311 COMPLAINTS
☐ Ice Cream truck
☐ Barking Dog
☐ Loud Music/party

APPLY

About • Help • Feedback

bit.ly/newerhoods
Evolution of Real Estate Market

5-yr (2013-17) average real estate prices

2017 average real estate prices

Yorkville got the Q train in 2017!
Violation Rates vs. Precincts

77 NewerHoods based on violation rates with no prior knowledge of precinct boundaries

NewerHoods mirror existing precincts - individual precinct behaviour shapes clusters
Misdemeanors vs. Precincts!?

77 NewerHoods based on misdemeanor rates with no prior knowledge of precinct boundaries

NewerHoods mirror precincts in some areas - precinct behaviour shouldn’t influence clusters for misdemeanors
311 Noise Complaints

- Loud music/parties
- Ice cream trucks
- Barking dogs

DATA CLINIC
Applications

→ Aid social-science research
  ◆ Definitions of neighborhoods critical to social-science researchers looking to model local-effects

→ Individual use
  ◆ Find your neighborhood based on personal preferences
  ◆ Download neighborhood report and summary

→ Civic tech
  ◆ Observe changing boundaries over time to detect/predict characteristics (i.e., gentrification)
Applications

➔ Aid social-science research
   ◆ Local/neighborhood effects important in predicting social and economic outcomes

➔ Civic tech
   ◆ Analysing changing boundaries over time could help predict things such as gentrification and aid city planning

➔ Individual use
   ◆ Neighborhood reports for community organizers
How you can contribute?

➔ Submit data sets to integrate into NewerHoods
➔ Help with our development efforts
  ◆ Upload your own data
  ◆ Extracting features from different geographic aggregations
  ◆ Improve clustering methodology
➔ Submit issues on our GitHub
➔ Feedback
  ◆ Tell us how you’d like to use NewerHoods in your work

DATA CLINIC

bit.ly/newerhoods
tsdataclinic/newerhoods
dataclinic@twosigma.com
Summary

➔ Making data more visual and accessible
➔ Laying the foundation for more development
➔ Open-sourced code
➔ What are we working on next
  ◆ Upload your own data
  ◆ Extracting features from different geographic aggregations
  ◆ Integrating additional data sets
<table>
<thead>
<tr>
<th>Which datasets should we use?</th>
<th>What kinds of features should we include?</th>
<th>How can we improve the methodology?</th>
</tr>
</thead>
<tbody>
<tr>
<td>How do people use different definitions of neighborhoods?</td>
<td>How would people use this tool?</td>
<td>How do we build community engagement?</td>
</tr>
</tbody>
</table>
Thank you

Darren Erik Vengroff
dataclinic@twosigma.com
Rachael Weiss Riley  ||  Erin Stein  ||  Kaushik Mohan
bit.ly/newerhoods
https://github.com/tsdataclinic/newerhoods