

Transforming Open Data into Insights

*How Data Clinic Uses Open Data to Support
Mission-Driven Organizations*

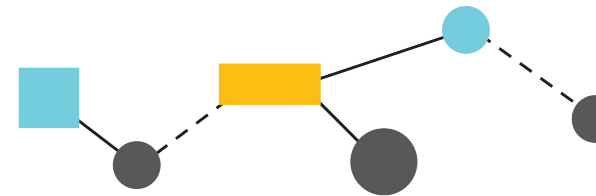
Erin Stein || [@tsdataclinic](#)



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





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-driven research and tooling to contribute to the data-for-good movement

How we work



engagement + team



research + development



results + impact

Data science projects



Can we match participants to training programs for maximize success?



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



Can we identify why some projects are more likely to be funded than others?

Common threads

- Established organizations
- ~~→ A lot of data in-house~~
- ~~→ Research questions that could be answered by in-house data~~

**Open
data!**

Why use open data?

- It exists!
- Open data is diverse
- Varied applications/use cases
 - ◆ Build business case for data strategy
 - ◆ Advance research



Why use open data?

Build a business case for
data strategy



Finding the ways that work

*What predicts future oil and gas
industry violations?*

Advance research

Vera

INSTITUTE OF JUSTICE


*Can we provide insight into the
national landscape of open 911 call
data?*

Building a business case for a data strategy

Build a business case for
data strategy



*What predicts future oil and gas
industry violations?*

-  **pennsylvania**
DEPARTMENT OF ENVIRONMENTAL
PROTECTION
- Past violations + inspection frequency
were highly predictive of future
violations
- Resulted in:
 - ◆ Culture shift at EDF
 - ◆ Shared, inter-organizational research
strategy

Advancing research

- Engaging in research to better understand 911 calls and police enforcement
- No nationally standardized data exists on 911 calls
 - ◆ Difficult to understand call volumes, what police are responding to, etc.
- Initial analysis on Seattle, New Orleans, Charleston, Dallas, and Detroit
 - ◆ Develop a pipeline to acquire, clean and standardize open 911 call data

Advance research

Vera

INSTITUTE OF JUSTICE

Can we provide insight into the national landscape of open 911 call data?

YARRRRRRRGHHH

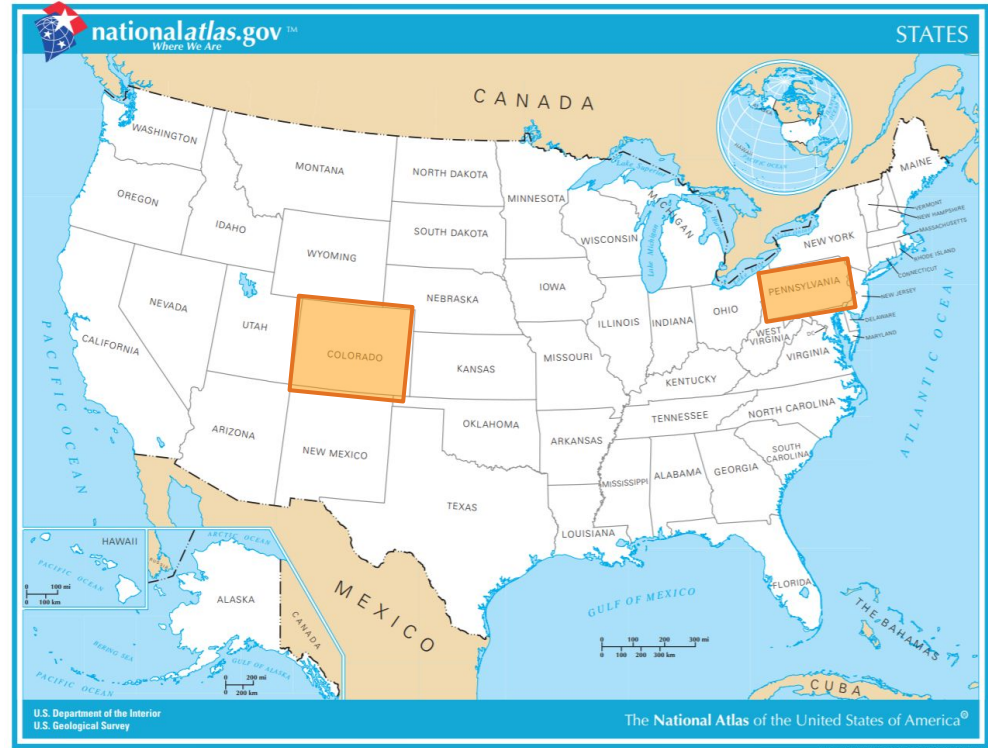


Nonexistent and unstructured data

Build a business case for
data strategy



**Only ONE of the two states with
open data on oil and gas
inspections had usable data!**



Inconsistencies across sources

| City | CFS Code | Call Type | Disposition | Lat-Long | Priority | Year Range | Beat/District |
|-------------|----------|-----------|-------------|----------|----------|------------|---------------|
| Charleston | Yes | No | Yes | Yes | No | '15 - '17 | No |
| Dallas | Yes | No | Yes | Yes | Yes | '05 - '19 | Yes |
| Detroit | Yes | Yes | No | Yes | Yes | '17 - '18 | Yes |
| New Orleans | Yes | Yes | Yes | Yes | Yes | '11 - '19 | Yes |
| Seattle | Yes | Yes | Yes | No | Yes | '09 - '19 | Yes |

Advance research

Vera

INSTITUTE OF JUSTICE

While all 5 cities had open 911 data, variables of interest and time spans were inconsistent!

Categorical minutiae

Advance research

Vera

INSTITUTE OF JUSTICE

Categories in the data didn't fit Vera's needs—they were either much too broad or excessively detailed!

◆ 24 broad categories

Top 6 CFS Categories

Complaints/Environmental Conditions

Statuses

Accidents/Traffic Safety

Suspicion

Assist the Public

Alarms

◆ Each category can contain upwards of 100 different CFS codes for any given city

Complaints/Environmental Conditions

DISTURBANCE, MISCELLANEOUS/OTHER

--MISCHIEF OR NUISANCE - GENERAL

HAZ - POTENTIAL THRT TO PHYS SAFETY (NO HAZMAT)

NOISE - DIST, GENERAL (CONST, RESID, BALL PLAY)

LOST PROPERTY

VICIOUS ANIMAL

FIREWORKS - NUISANCE (NO HAZARD)

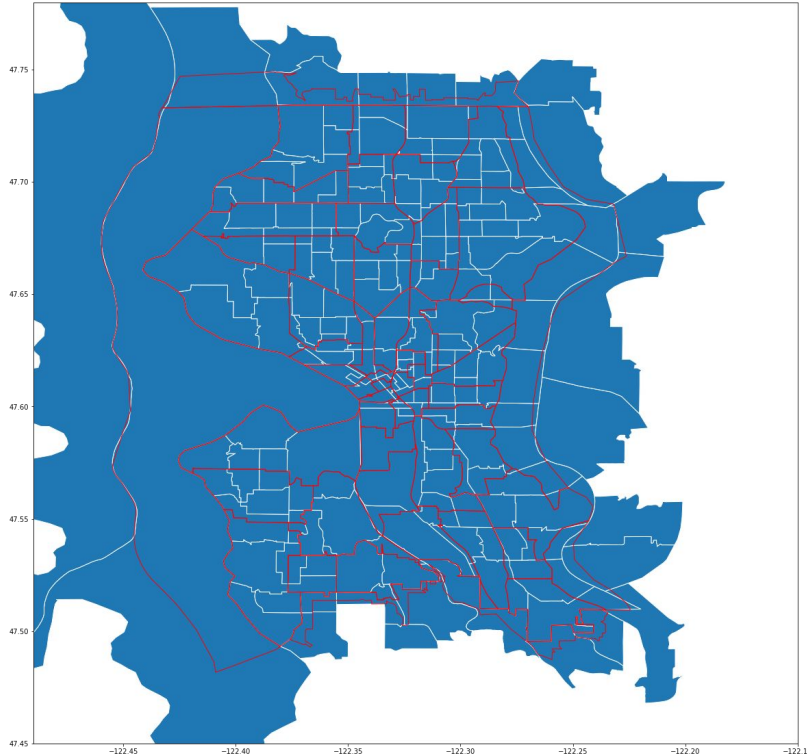
QUALITY OF LIFE ISSUE

--ANIMAL COMPLAINT - NOISE,STRAY,BITE

HOMELESS

SQUATTER DISTURBANCE

Geospatial inconsistencies



Advance research

Vera INSTITUTE
OF JUSTICE

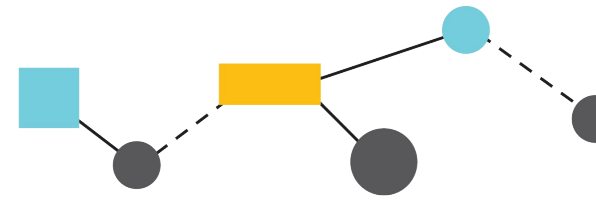
**Geolocation data varied across
datasets and cities!**

Open data is HARD



Vera
INSTITUTE OF JUSTICE

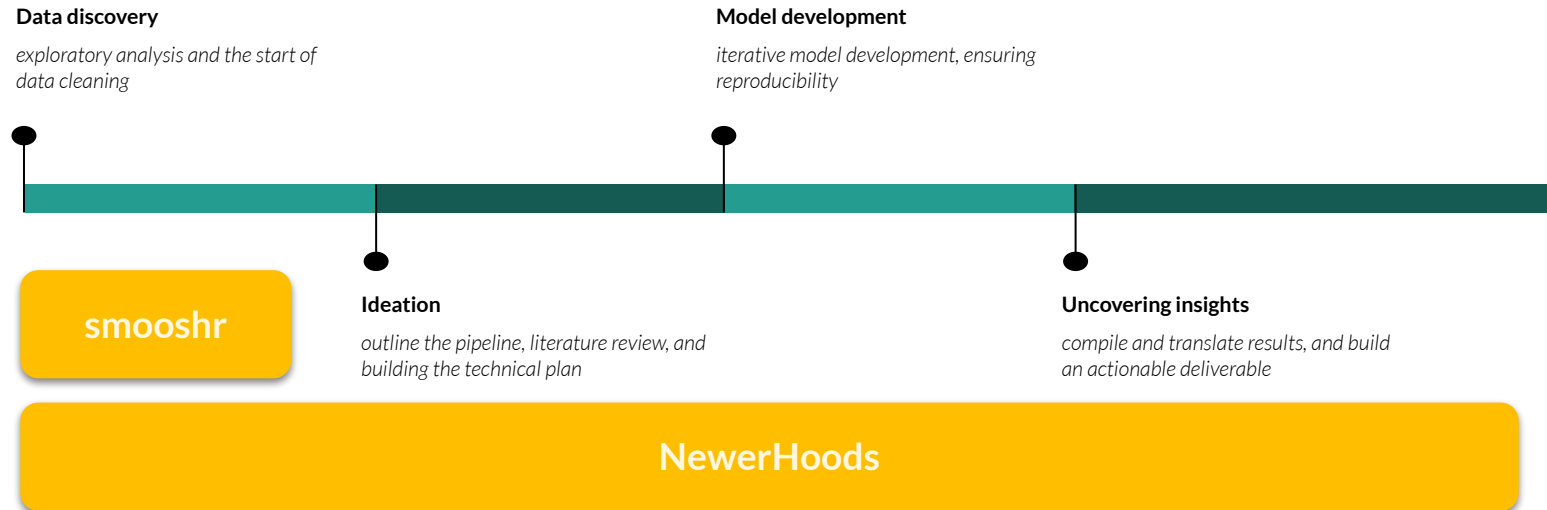
- Open data is incomplete
- Open data is messy
 - ◆ A lot of free-form text fields
 - ◆ Lack of standards in data entry
 - ◆ Changing variable names over time
- Original purpose of data collection may not match purpose of the research
- Just because it's open, doesn't mean it's accessible



a little bit of tooling could go a long way

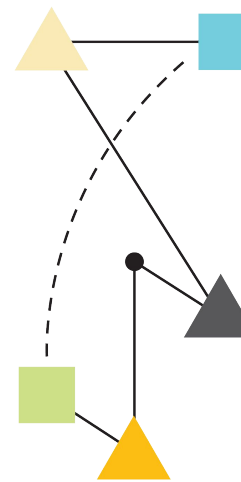
Data science pipeline

→ Developing open source tools throughout the data science pipeline



smooshr

Facilitating entity consolidation of messy data



Creating meaningful variables and taxonomies

Animals

| Complaints/Environmental Conditions |
|---|
| DISTURBANCE, MISCELLANEOUS/OTHER |
| --MISCHIEF OR NUISANCE - GENERAL |
| HAZ - POTENTIAL THRT TO PHYS SAFETY (NO HAZMAT) |
| NOISE - DIST, GENERAL (CONST, RESID, BALL PLAY) |
| LOST PROPERTY |
| VICIOUS ANIMAL |
| FIREWORKS - NUISANCE (NO HAZARD) |
| QUALITY OF LIFE ISSUE |
| --ANIMAL COMPLAINT - NOISE, STRAY, BITE |
| HOMELESS |
| SQUATTER DISTURBANCE |

- Understand and join non-uniform, messy text data
 - ◆ Create appropriate aggregations
 - ◆ Consolidate columns across years or sources that reference same variable
 - ◆ Build standard taxonomy within consolidated columns
- Building a tool to facilitate this process through:
 - ◆ A user-friendly UI
 - ◆ ML approaches to variable category suggestions

smooshr || 1. create a project

→ Projects organize all work

- ◆ They contain datasets
- ◆ And the taxonomies you create for them

The screenshot shows the 'smooshr' web application interface. The header includes the 'smooshr' logo and the tagline 'wrangle those messy datasets'. A 'New Project' button is visible in the top right corner. The main content area is divided into two sections: 'your projects' and 'community projects'. Under 'your projects', there are two project cards. The first card is titled 'New Orleans Calls for Action' and describes 'multiple years of 311 complaint types from 2017-2019' with a date range of '20' to '2' and a count of '300000'. The second card is titled '311 Complaints in New York and Chicago' and describes 'combining data from chicago and new york to look at trends in broken windows policies' with the same date range and count. The 'community projects' section is currently empty. A left sidebar contains the 'smooshr' logo and a navigation arrow.

smooshr || 2. load in datasets

- Datasets can be loaded from 3 different sources
 - ◆ local files
 - ◆ urls
 - ◆ directly from Socrata open data portals
- Currently we only support tabular data but aim to expand in future



smooshr || 3. group and rename columns

- Columns from different datasets are often measuring similar things in different circumstances
 - ◆ Ex: 911 call reasons across years
- smoosher lets you collapse these columns into a new column and generate taxonomies for the combined dataset

The screenshot displays the Smooshr interface for a project titled "311 Complaints in New York and Chicago". The interface is organized into several sections:

- Datasets:** Shows two source datasets: "nola_calls_for_service_2017.csv" and "nola_calls_for_service_2018.csv".
- Columns:** Displays a grid of columns being mapped from the source datasets to a new column.

| Source Column | Count | Target Column | Count |
|-----------------------|-------|---------------------|-------|
| InitialTypeText (166) | 166 | Disposition (9) | 9 |
| InitialTypeText (170) | 170 | DispositionText (9) | 9 |
- Actions:** Provides buttons for "Export Mappings (csv)", "Export Mappings (json)", "Export Python code", and "Export Data".

smooshr || 3. group and rename columns

- Columns from different datasets are often measuring similar things in different circumstances
 - ◆ Ex: 911 call reasons across years
- smoosher lets you collapse these columns into a new column and generate taxonomies for the combined dataset

311 Complaints in New York and Chicago
combining data from chicago and new york to look at trends in broken windows policies

Datasets
nola_calls_for_service_2017.csv nola_calls_for_service_2018.csv

Columns

| | | | |
|--|-----|---|---|
| Disposition <input type="checkbox"/> | 9 | DispositionText <input type="checkbox"/> | 9 |
| disposition nola_calls_for_service_2017.csv work on mappings | | dispositiontext nola_calls_for_service_2018.csv work on mappings | |
| InitialTypeText <input type="checkbox"/> | 336 | | |
| initialtypetext nola_calls_for_service_2017.csv initialtypetext nola_calls_for_service_2018.csv work on mappings | | | |

Actions

Export Mappings (csv) Export Mappings (json) Export Python code Export Data

smooshr || 4. create taxonomies for each column

- ➔ Search for unique categories in the combined columns
 - ◆ Group multiple entries into new taxonomies
 - ◆ New taxonomy can be renamed
- ➔ smooshr sends individual words that make up an entry to a server to get word embeddings
 - ◆ Suggests other entries that might belong to the current taxonomy

The screenshot displays the smooshr interface with a search bar at the top containing the text 'Q'. Below the search bar, the results are ordered by 'Occurrences'. The interface shows a grid of 18 incident categories, each with a name and a count of occurrences:

| Category | Occurrences |
|--------------------------------------|-------------|
| complaint other | 136290 |
| disturbance (other) | 52899 |
| suspicious person | 34956 |
| theft | 26600 |
| hit & run | 15848 |
| simple battery | 10959 |
| auto accident with injury | 9505 |
| area check | 71845 |
| auto accident | 38518 |
| return for additional info | 29759 |
| domestic disturbance | 18068 |
| fight | 13044 |
| directed patrol | 10236 |
| simple criminal damage | 9233 |
| burglar alarm, silent | 59402 |
| warr stop with release | 36330 |
| business check | 29384 |
| municipal attachment | 15905 |
| auto theft | 12071 |
| incident requested by another agency | 10379 |
| fugitive attachment | 8937 |

At the bottom of the interface, there is a 'MERGED GROUPS' section showing 'traffic incident' with 1 entry. To the right, a 'TRAFFIC INCIDENT' section shows a list of suggestions by text / meaning:

| Suggestion | Status |
|--------------------------------------|-------------------------------------|
| traffic congestion | <input checked="" type="checkbox"/> |
| traffic attachment | <input checked="" type="checkbox"/> |
| auto accident police vehicle | <input checked="" type="checkbox"/> |
| incident requested by another agency | <input checked="" type="checkbox"/> |
| police vehicle accident | <input checked="" type="checkbox"/> |
| directed traffic enforcement | <input checked="" type="checkbox"/> |
| auto accident fatality | <input checked="" type="checkbox"/> |
| auto accident city vehicle | <input checked="" type="checkbox"/> |

Buttons for 'New Mapping 0', 'Add to Mapping', 'Map Remaining To Other', 'Clear Mapping', and 'Delete Mapping' are also visible.

smooshr || 4. create taxonomies for each column

- Search for unique categories in the combined columns
 - ◆ Group multiple entries into new taxonomies
 - ◆ New taxonomy can be renamed
- smooshr sends individual words that make up an entry to a server to get word embeddings
 - ◆ Suggests other entries that might belong to the current taxonomy

The screenshot shows the smooshr interface. At the top, there is a search bar with the text 'theft' and a search icon. Below the search bar, the results are displayed as a grid of boxes, each representing a taxonomy. The boxes are:

- complaint other (136290 occurrences)
- disturbance (other) (52899 occurrences)
- theft from exterior (2201 occurrences)
- bicycle theft (312 occurrences)
- theft from exterior of vehicle (85 occurrences)

At the bottom of the interface, there is a 'MERGED GROUPS' section with 2 groups. The first group is 'THEFT' and it includes the following terms:

- theft
- theft by fraud
- business burglary
- simple burglary
- theft by embezzlement
- auto theft
- simple burglary vehicle
- safe burglary

Below the 'THEFT' group, there are 'Suggestions by: text / meaning' which include:

- simple burglary domestic
- residence burglary
- theft from exterior
- aggravated burglary domestic
- bicycle theft
- aggravated burglary
- theft from exterior of vehicle
- simple robbery

At the bottom of the interface, there are buttons for 'Clear Mapping' and 'Delete Mapping'.

smooshr || 5. export the results

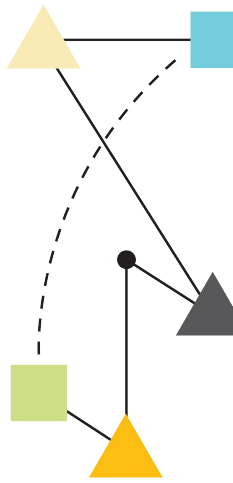
→ Results / mappings can be exported

- ◆ csv file (in development)
- ◆ JSON document
- ◆ Python code snippet
 - Can be applied to the files in an ETL workflow
- ◆ The transformed data!

The screenshot displays the 'smooshr' application interface. On the left, a teal sidebar contains the 'smooshr' logo and a back arrow. The main content area is titled '311 Con' and shows a 'Datasets' section with a file named 'nola_calls_for_service_2017.csv'. Below this, the 'Columns' section lists 'Disposition' and 'InitialTypeText' with checkboxes. At the bottom, an 'Actions' section contains four buttons: 'Export Mappings (csv)', 'Export Mappings (json)', 'Export Python code', and 'Export Data'. A 'Save File' dialog box is open in the foreground, showing the file name '311-Complaints-in-New-York-and-Chicago.json' and a file explorer view of the 'stuart tmp Mappings' directory. The dialog has 'Cancel' and 'Save' buttons at the bottom right.

NewerHoods

uncovering patterns in challenging geospatial data



Neighborhoods are newsworthy

Long Island City — future home to Amazon HQ — is one of NYC’s hottest new spots for young people

Laura Begley Bloom, Special to CNBC | 10:30 AM ET Sat, 17 Nov 2018



Report Reveals NYC Neighborhoods with Highest Asthma Rates

By: Rob Senior
February 5, 2019

Share: [f](#) [in](#) [t](#) [g+](#)

NYC AFFORDABLE HOUSING NEWS

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

[f](#) [t](#) [SHARE](#)



New York City’s Biggest ‘Food Swamps’

By Lea Ceasrine | May 21, 2018

[f](#) RECOMMEND [t](#) TWEET [e](#) EMAIL [p](#) PRINT [+](#) MORE

Central Brooklyn. Bronx notorious for

website indicates the Brownsville section has the city.

lize.city comes on the heels of the Asthma Free on January 19, and aims to protect tenants from beaches, mice, and rats.

What is a neighborhood?

Social

common perceptions

- No clear definition
- Varies from person to person
- Eg: “hip” neighborhoods in Brooklyn

Administrative

who is served

- Defined specifically to serve the respective organization
- Optimized based on organizational costs
- Static definitions
- Eg: police precincts or ZIP codes

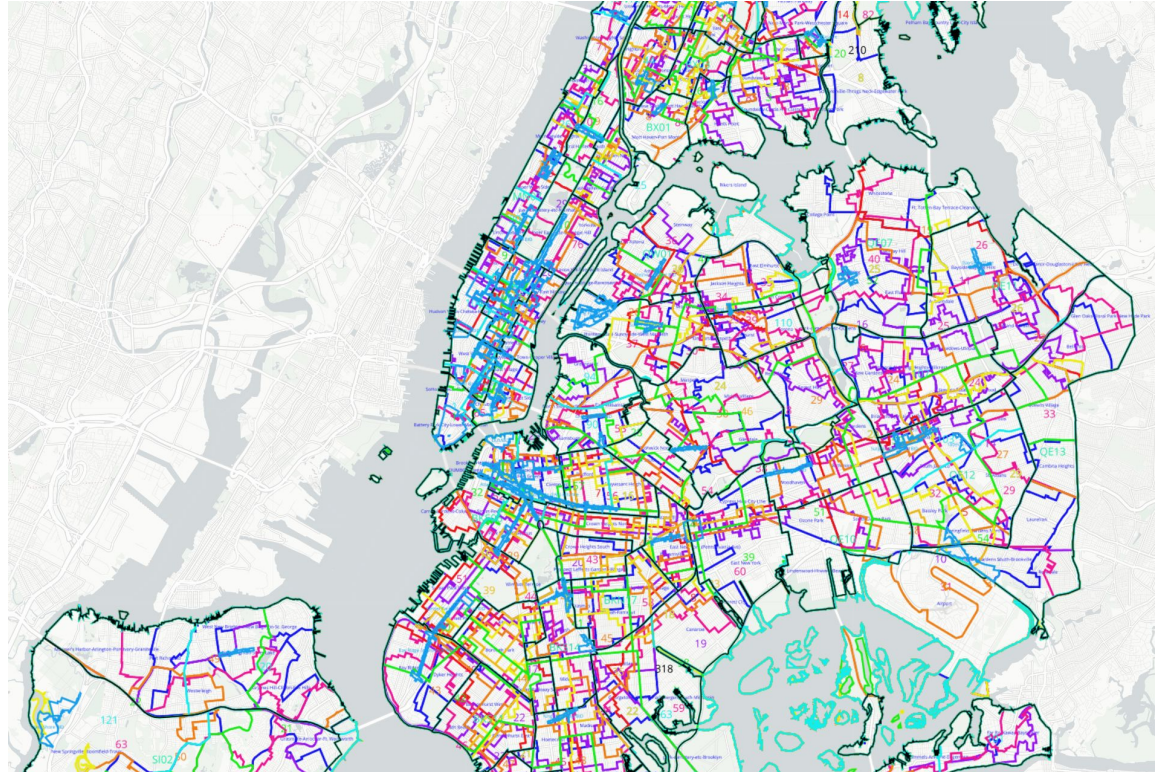
Statistical

data collection

- Defined to capture areas with a specific population count for data collection & organization
- Updated approximately every 10 years
- Eg: US Census block groups or tracts

Why do we care?

- Community Districts —
- Police Precincts —
- Sanitation Districts —
- Fire Battalions —
- School Districts —
- Health Center Districts —
- City Council Districts —
- Congressional Districts —
- State Assembly Districts —
- State Senate Districts —
- Neighborhood Tabulation Areas —
- Business Improvement District —



Source: Beta.NYC

Some current tools

MAKING NEIGHBORHOODS (Citizens Housing and Planning Council)

Cluster Name
Majority white, low-middle income, singles and couples

Race Composition

| | |
|-----------|-----|
| Asian: | 7% |
| Black: | 12% |
| Hispanic: | 23% |
| Whites: | 55% |
| Other: | 2% |

Age Composition

| | |
|--------|-----|
| 0-18: | 21% |
| 18-34: | 25% |
| 35-64: | 41% |
| 65+: | 14% |

Family Composition

| | |
|----------------------|-----|
| One Person: | 31% |
| Shared Non-Family: | 7% |
| Single Parents: | 8% |
| Couples w/Children: | 19% |
| Couples wo/Children: | 24% |
| Other wo/Children: | 11% |

Foreign 29%

(topos.com) **22 CLUSTERS** 33

topos

Discover Community Lifestyle and Demographic Information (Environmental Systems Research Institute (ESRI))

10024

Analyze: Drive Time Ring Buffer

North Bergen West New York Union City Astoria

Tapestry

| | |
|-------|--------------------|
| 87.2% | Laptops and Lattes |
| 6.5% | Urban Chic |
| 2.3% | Golden Years |

Segment Description
We are affluent, well-educated singles and partner couples who love life in the big city and hold professional positions. Most rent apartments and either work from home or walk, bike, and take.

Median Income

| |
|-----------|
| \$116,403 |
| \$77,442 |
| \$54,756 |

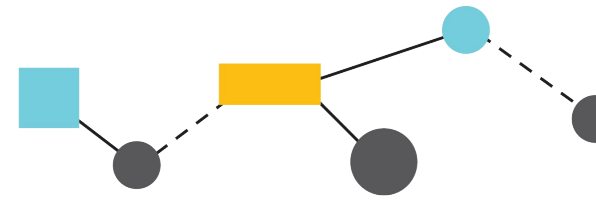
Graduate and Professional Degrees

| |
|-------|
| 43.8% |
| 28.8% |
| 13.5% |

Population Density

| |
|----------|
| 98,647.3 |
| 72,908.4 |
| 378 |

Read more



An open, flexible, and dynamic tool

The approach

Open Data

Gather information on a variety of dimensions from NYC Open Data

Local Attributes

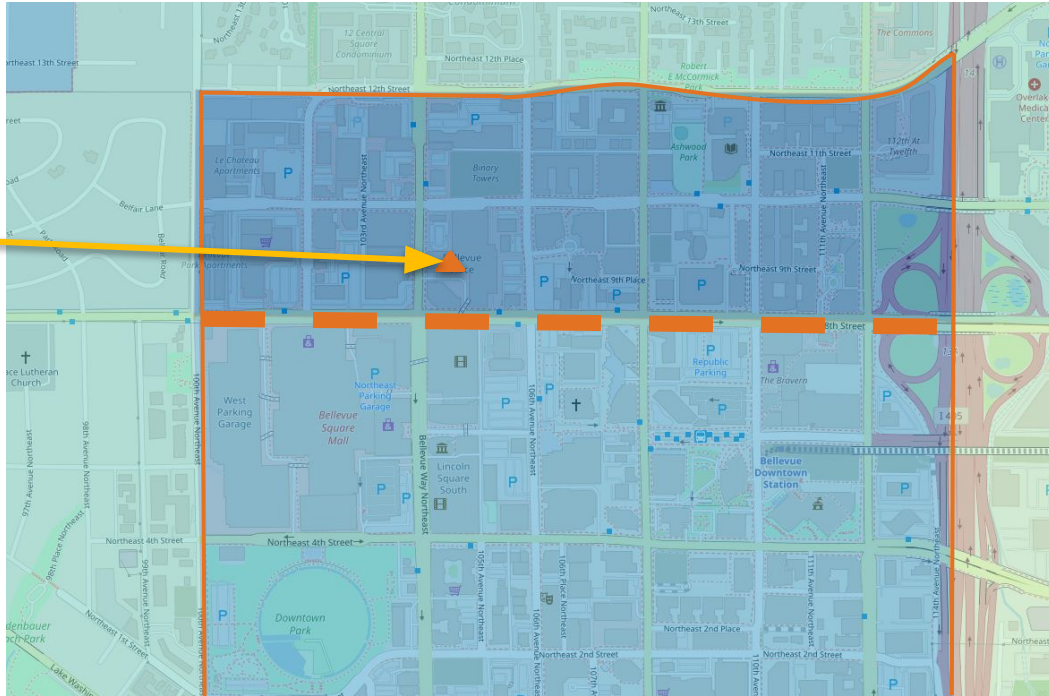
Extract multiple different attributes for every census tract from these data sets

Clustering

Use Machine Learning techniques to find homogenous areas based on chosen characteristics

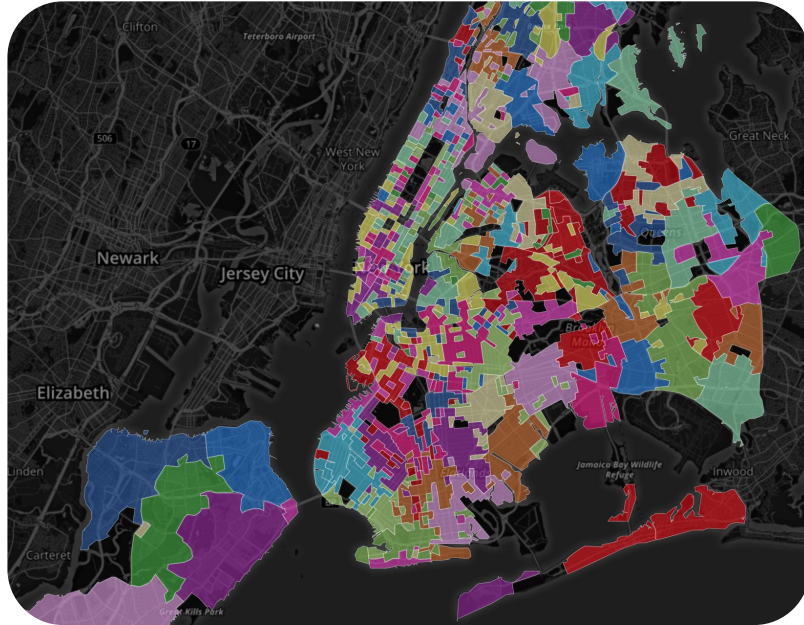
What is a census tract?

We are here
(900 Bellevue Way NE)



Traditional clustering methods fail

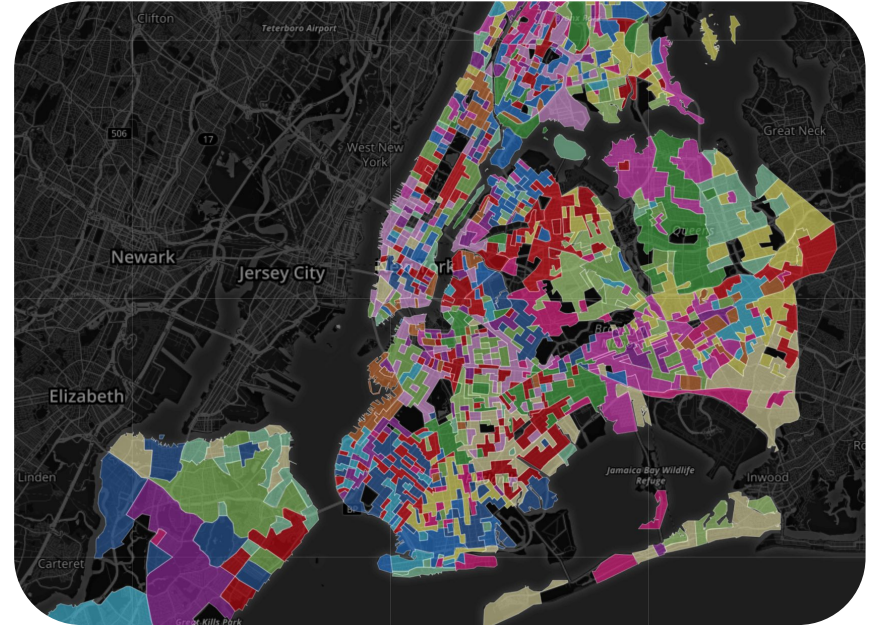
k-means clustering



Scaled features: Mean & sd of price per sq. footage, lat & lon of census tract (k = 100)

 TWO SIGMA

k-means clustering



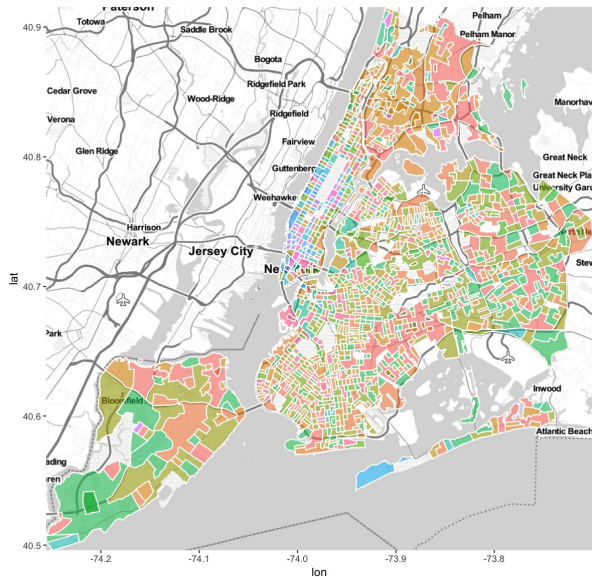
Scaled features: Mean & sd of price per sq. footage, violation rate, noise complaints, lat & lon of census tract (k = 100)

 data clinic

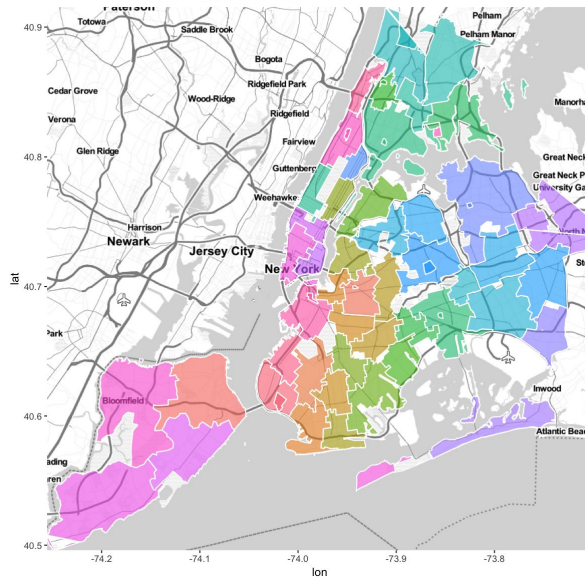
ClustGeo in action

$$I_{\alpha}(C_k^{\alpha}) = (1 - \alpha) \sum_{i \in C_k^{\alpha}} \sum_{j \in C_k^{\alpha}} \frac{w_i w_j}{2\mu_k^{\alpha}} d_{0,ij}^2 + \alpha \sum_{i \in C_k^{\alpha}} \sum_{j \in C_k^{\alpha}} \frac{w_i w_j}{2\mu_k^{\alpha}} d_{1,ij}^2,$$

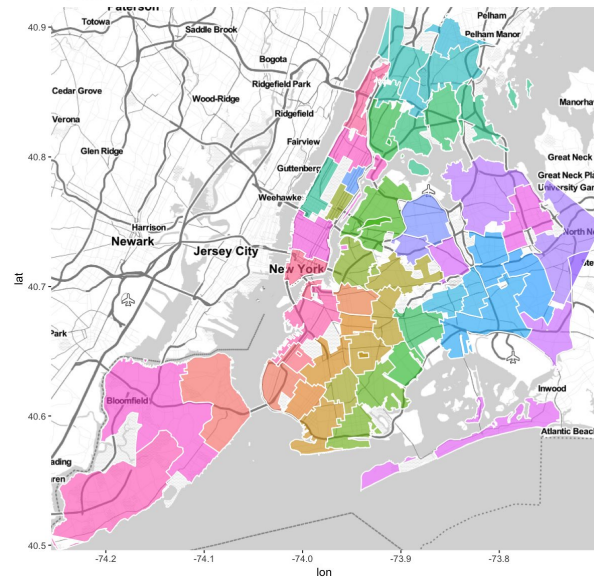
alpha= 0, k = 50



alpha= 0.15, k = 50



alpha= 1, k = 50



← Feature space: mean & sd of price/sq. ft

Geographic space: tract contiguity →

UPLOAD YOUR DATA

Choose characteristics to draw neighborhoods.

HOUSING

- Age of buildings
- Median sale price
 - 1y average
 - 3y average
 - 5y average

CRIME

- Violations
- Felonies
- Misdemeanors

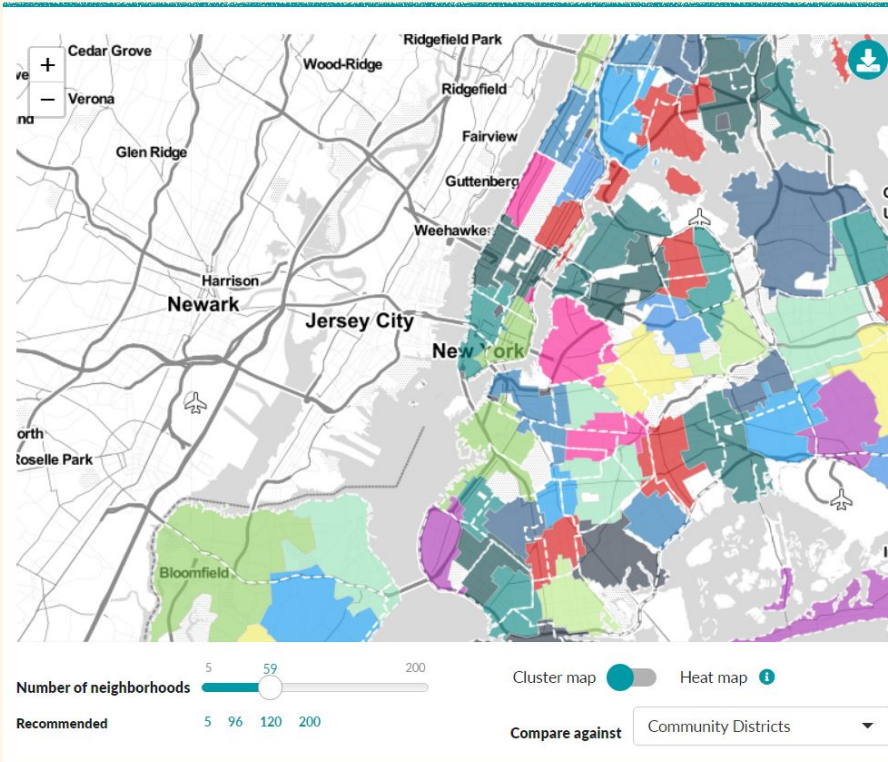
311 COMPLAINTS

- Noise related
- Housing related
- Street/Sidewalk related

APPLY

About • Help • Feedback

bit.ly/newerhoods



Applications

- Aid social-science research
 - ◆ Local/neighborhood effects important in predicting social and economic outcomes
- Civic tech
 - ◆ Analyzing changing boundaries over time could help predict things such as gentrification and aid city planning
- Individual use
 - ◆ Neighborhood reports for community organizers

Summary

- Pro bono data science rules!
- Use open data
 - ◆ Fills data gaps
 - ◆ Low/no-cost proof of concept
 - ◆ Expand current reach of research
- Build tools for repetitive tasks
 - ◆ Helps you AND helps others



Contribute

- Connect on GitHub
 - ◆ Take part in tool development
 - ◆ Submit issues
- Email us
 - ◆ Provide feedback
 - ◆ Refer potential use cases
- Visit our website
 - ◆ Follow our progress on projects and tooling



thank you

Erin Stein

dataclinic.twosigma.com

 [@tsdataclinic](https://twitter.com/tsdataclinic)

dataclinic@twosigma.com