data clinic

Transforming Open Data into Insights

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

Erin Stein || @tsdataclinic

TWO SIGMA

Disclaimer

TWO SIGMA

data clinic

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

THE HOPE PROGRAM
Work. Grow. Sustain.

Can we match participants to training programs for maximize success?

moulton niguel water district

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

DonorsChoose.org
Support a classroom. Build a future.

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

EDF ENVIRONMENTAL DEFENSE FUND®

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

EDF
ENVIRONMENTAL DEFENSE FUND®
Finding the ways that work

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?

YARRRRRRGGHHHH

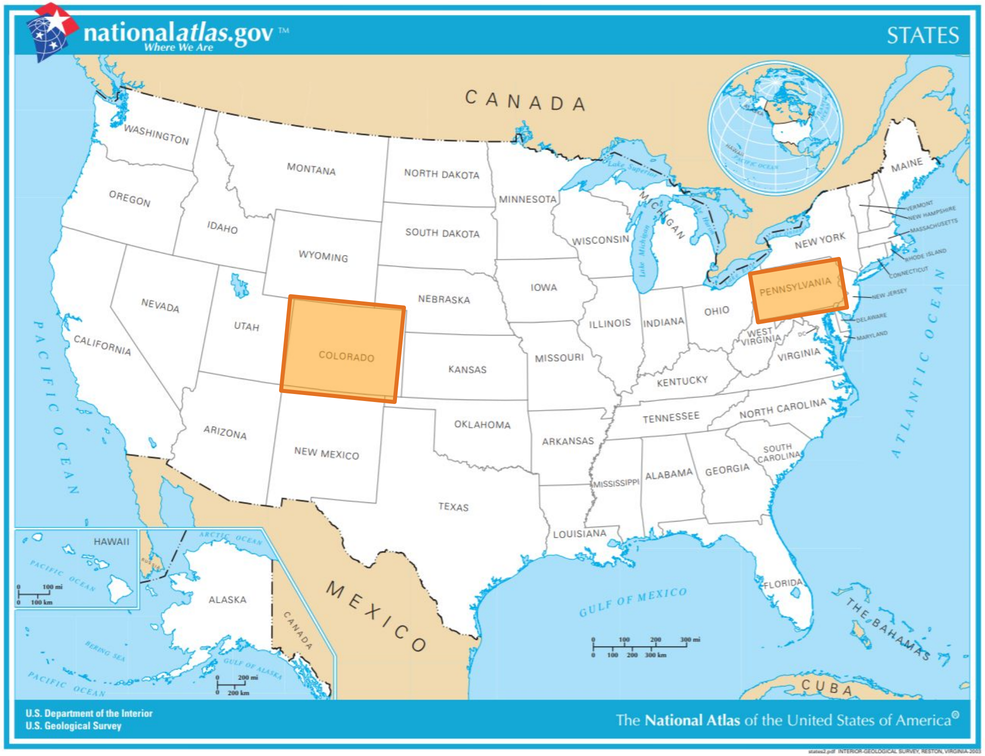
Nonexistent and unstructured data

Build a business case for data strategy

ENVIRONMENTAL DEFENSE FUND®

Finding the ways that work

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

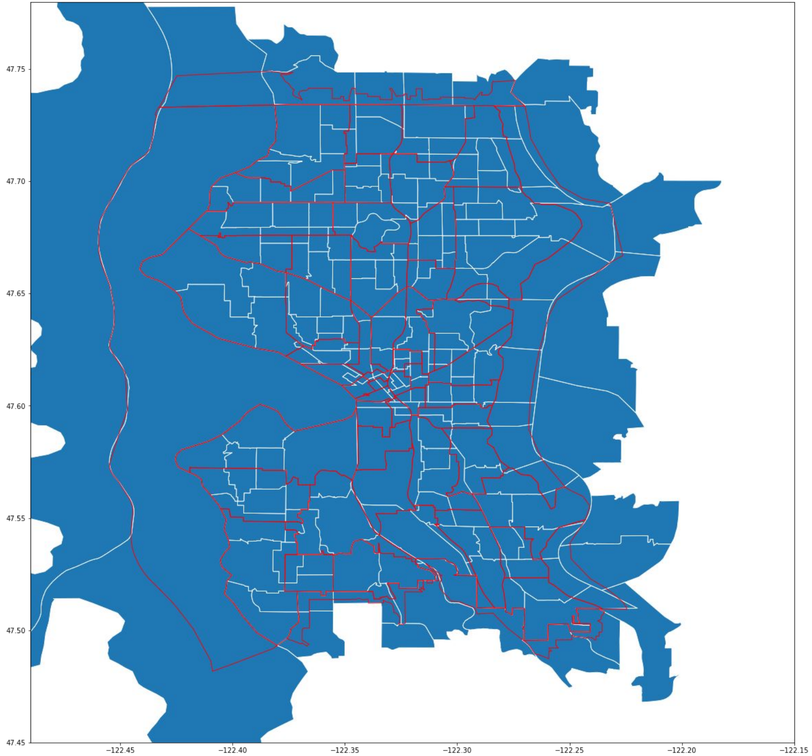
Categories in the data didn't fit Vera's needs—they were either much too broad or excessively detailed!

- 24 broad categories
- Each category can contain upwards of 100 different CFS codes for any given city

Top 6 CFS Categories
Complaints/Environmental Conditions
Statuses
Accidents/Traffic Safety
Suspicion
Assist the Public
Alarms

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

EDF
ENVIRONMENTAL DEFENSE FUND®
Finding the ways that work

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

Data discovery

exploratory analysis and the start of data cleaning

Ideation

outline the pipeline, literature review, and building the technical plan

Model development

iterative model development, ensuring reproducibility

Uncovering insights

compile and translate results, and build an actionable deliverable

smooshr

NewerHoods

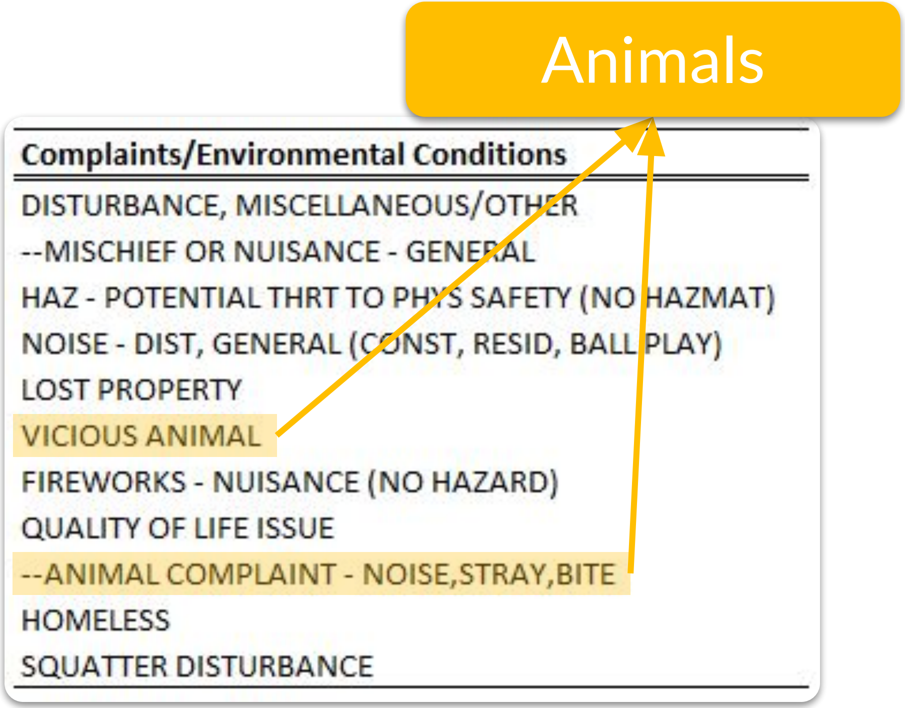
smooshr

Facilitating entity consolidation of messy data

TWO SIGMA

data clinic

Creating meaningful variables and taxonomies

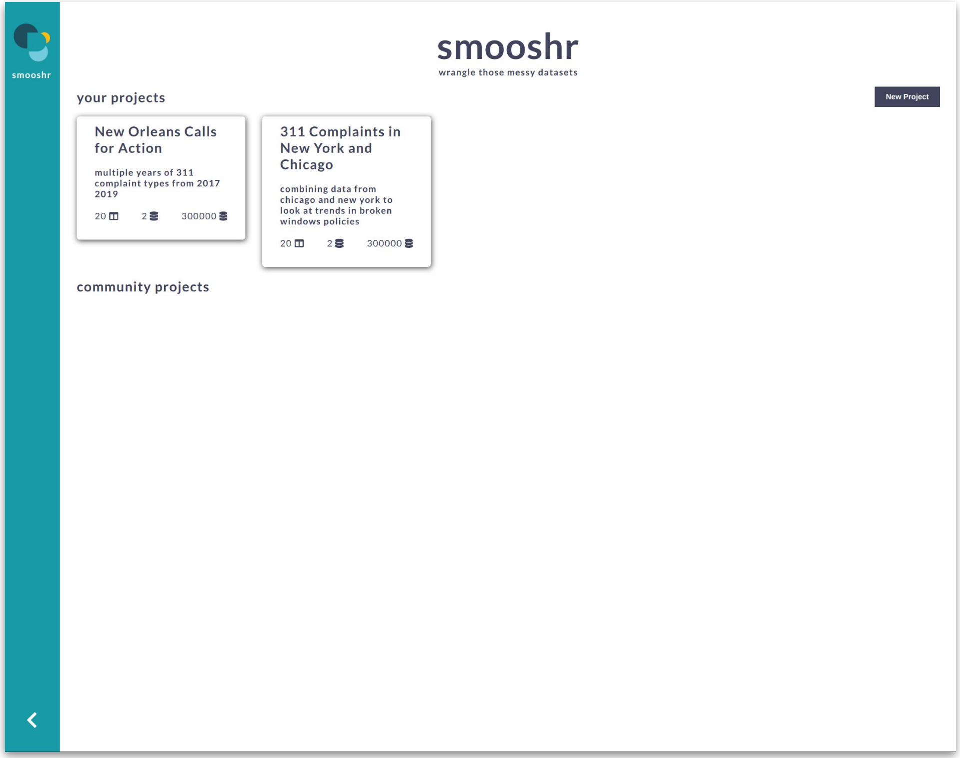


Complaints/Environmental Conditions
DISTURBANCE, MISCELLANEOUS/OTHER
--MISCHIEF OR NUISANCE - GENERAL
HAZ - POTENTIAL THRT TO PHYS SAFETY (NO HAZMAT)
NOISE - DIST, GENERAL (CONST, RESID, BALL PLAY)
LOST PROPERTY
VICIOUS ANIMAL
FIREWORKS - NUISANCE (NO HAZARD)
QUALITY OF LIFE ISSUE
--ANIMAL COMPLAINT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

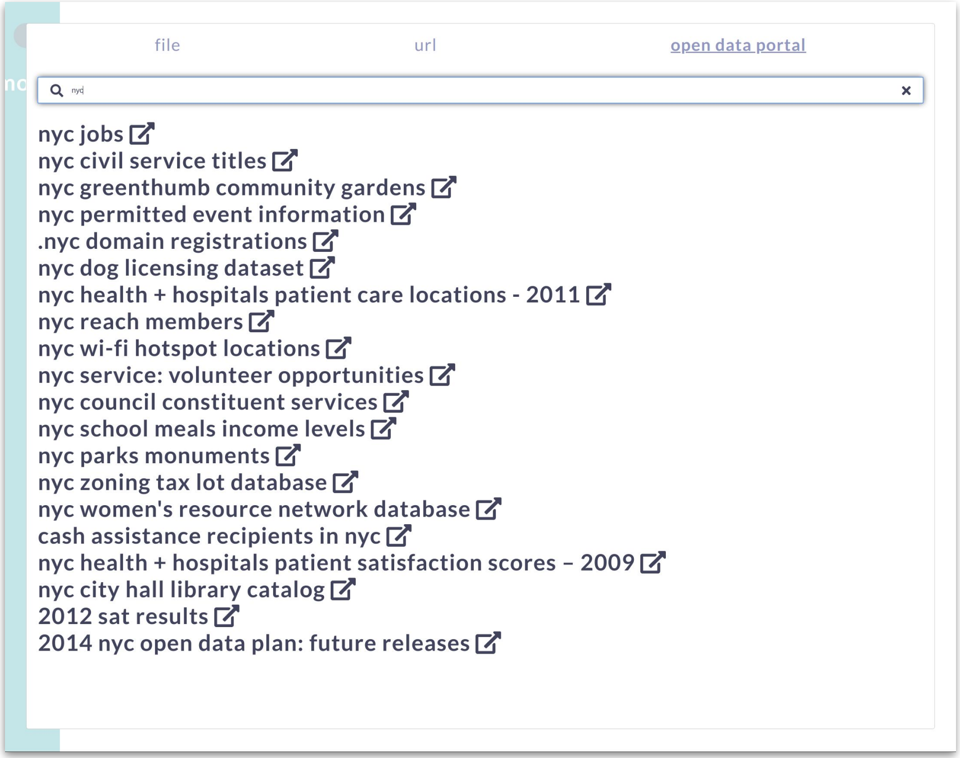
smooshr || 1. create a project

- Projects organize all work
  - They contain datasets
  - *And* the taxonomies you create for them



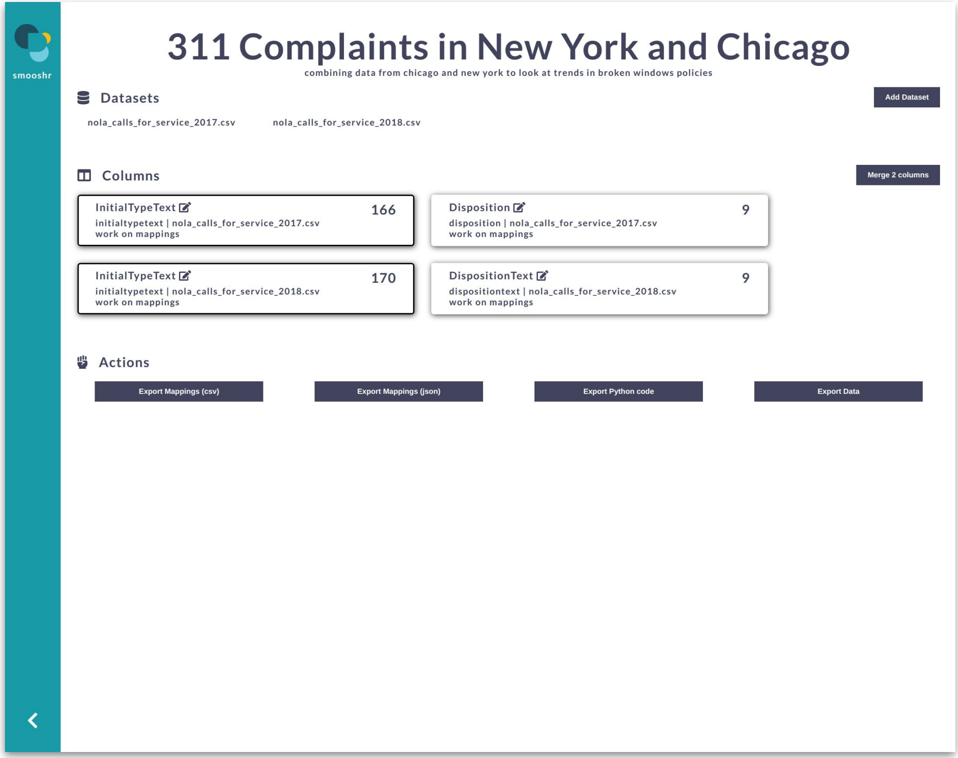
smooshr || 2. load in datasets

- 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

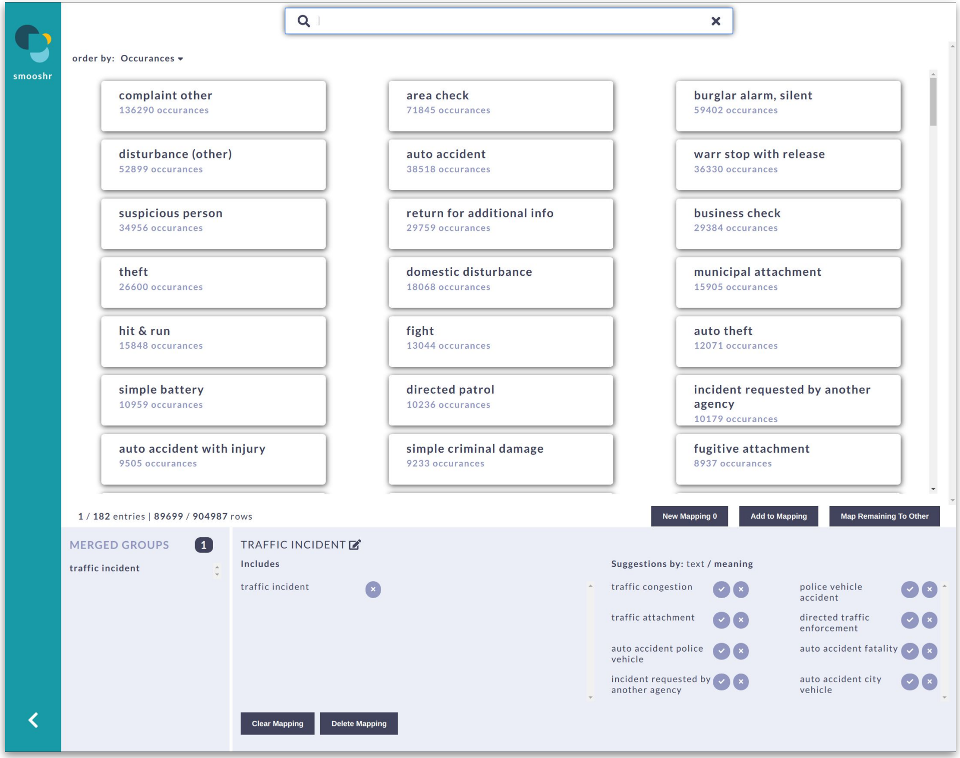


smooshr || 3. group and rename columns

- 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

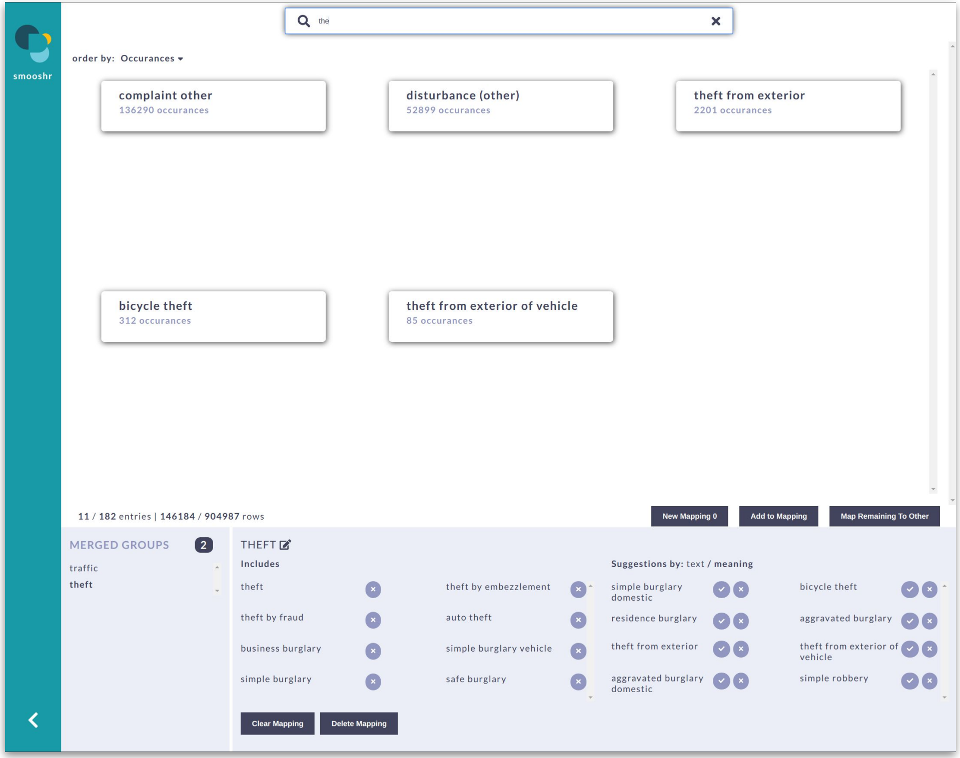
smooshr || 4. create taxonomies for each column

- → 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



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



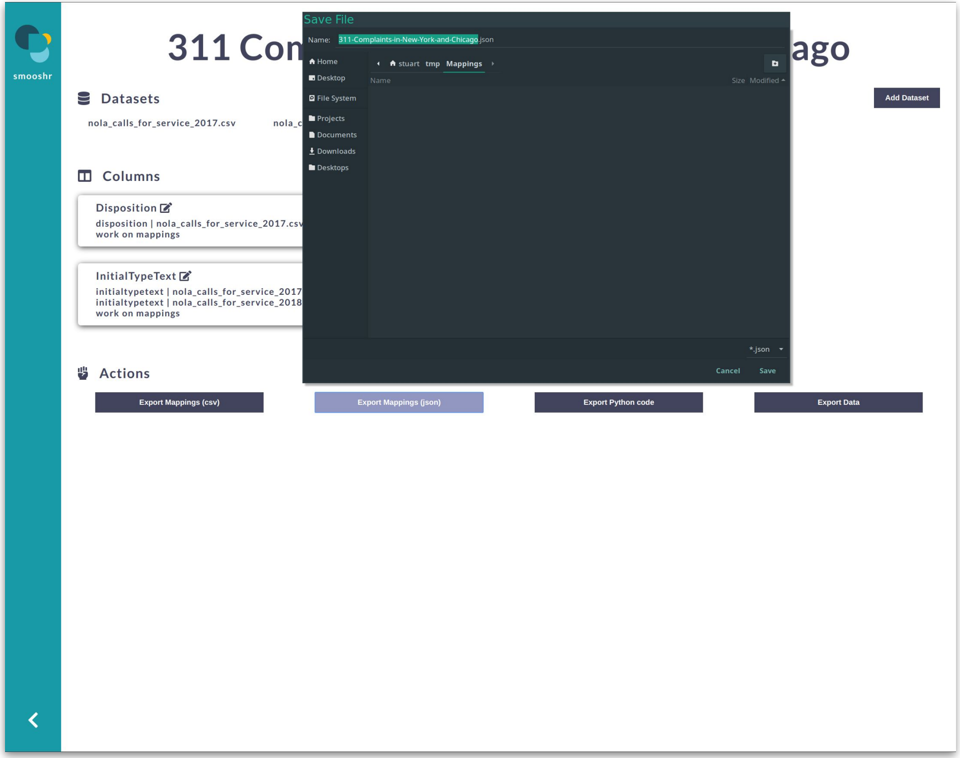
TWO SIGMA

data clinic

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!



NewerHoods

uncovering patterns in challenging geospatial data

TWO SIGMA

data clinic

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:

ntral Brooklyn, Bronx notorious for

website indicates the Brownsville section has the
ity.

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

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

SHARE

New York City's Biggest 'Food Swamps'

By **Lea Ceasrine** | May 21, 2018

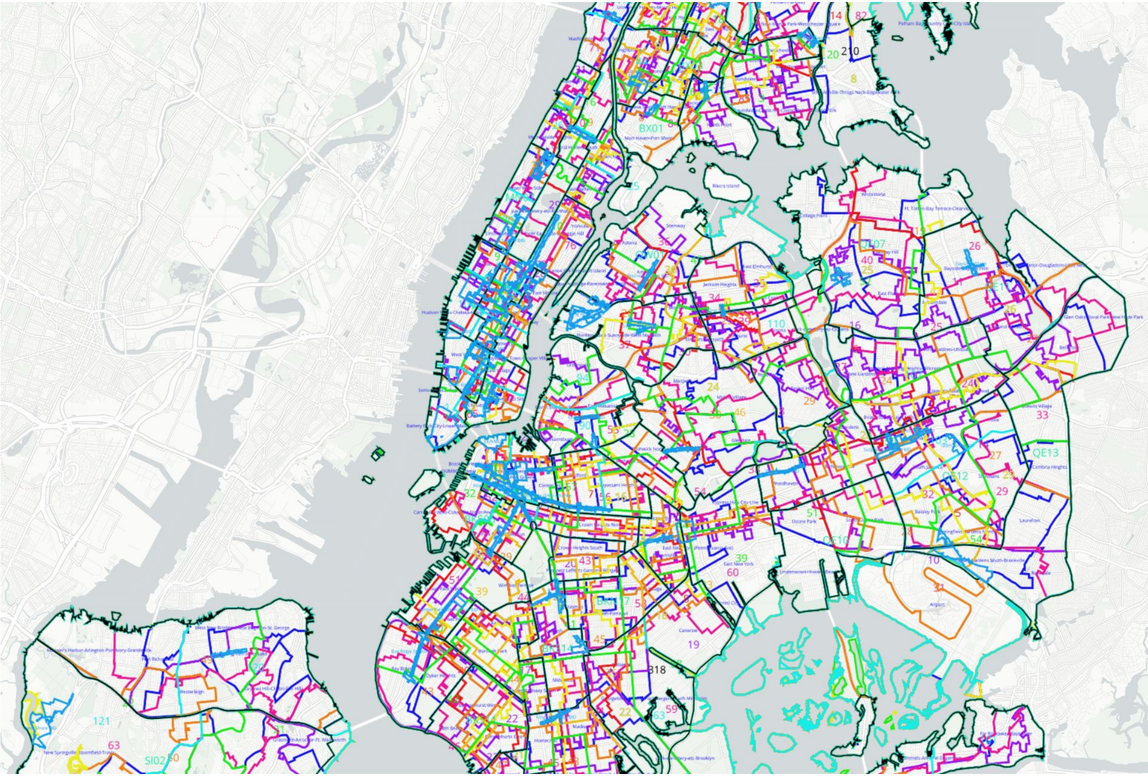
RECOMMEND | TWEET | EMAIL | PRINT | MORE

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

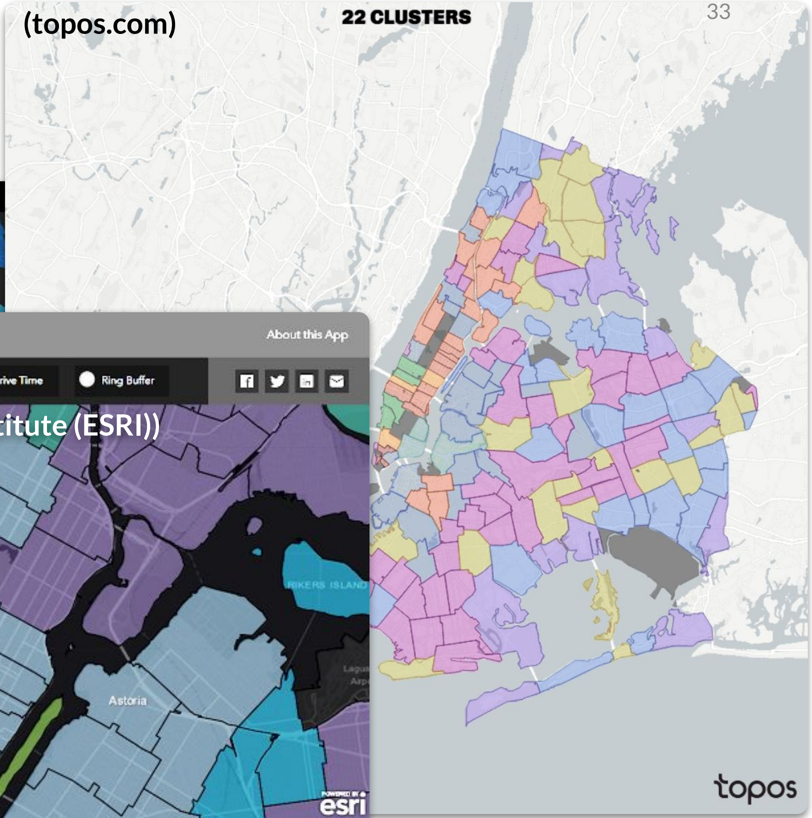
Why do we care?

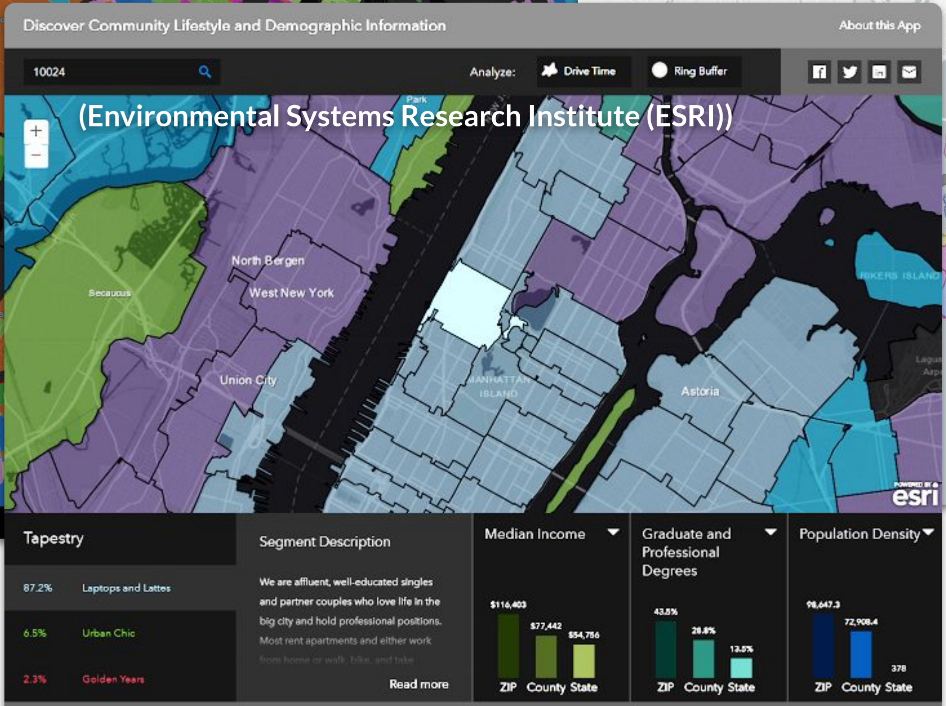
- Community Districts
- Police Precincts
- Sanitation Districts
- Fire Battilions
- School Districts
- Health Center Districts
- City Council Districts
- Congressional Districts
- State Assembly Districts
- State Senate Districts
- Neighbhorhood Tabluation Areas
- Business Improvement District



Source: Beta.NYC

Some current tools





An open, flexible, and dynamic tool

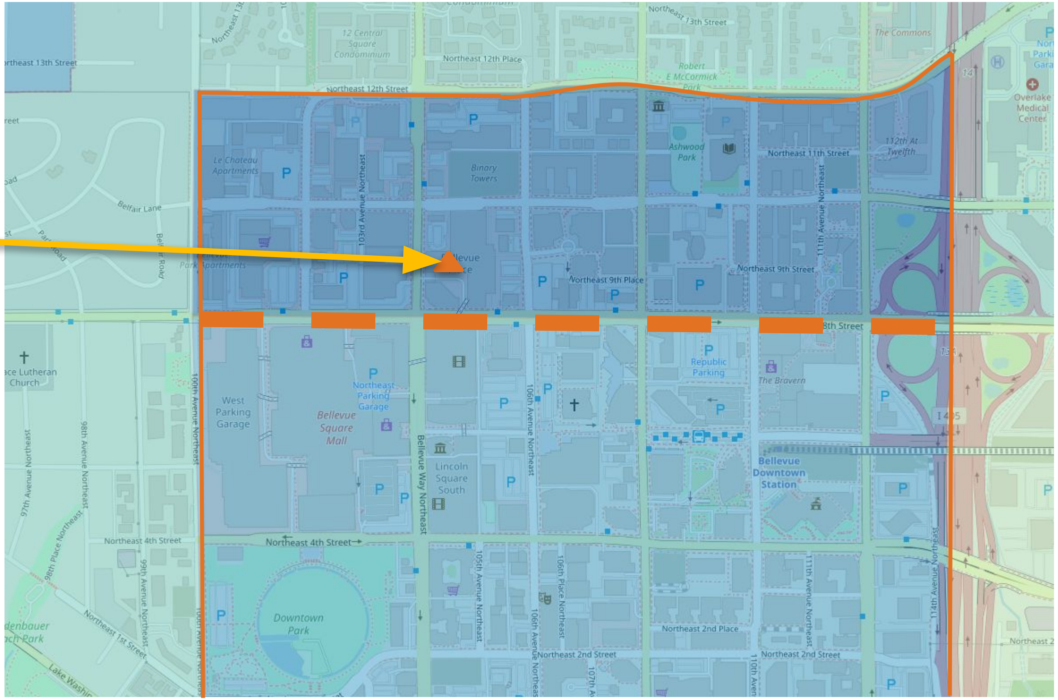
The approach

Open Data	Local Attributes	Clustering
Gather information on a variety of dimensions from NYC Open Data	Extract multiple different attributes for every census tract from these data sets	Use Machine Learning techniques to find homogenous areas based on chosen characteristics

What is a census tract?

We are here

(900 Bellevue Way NE)



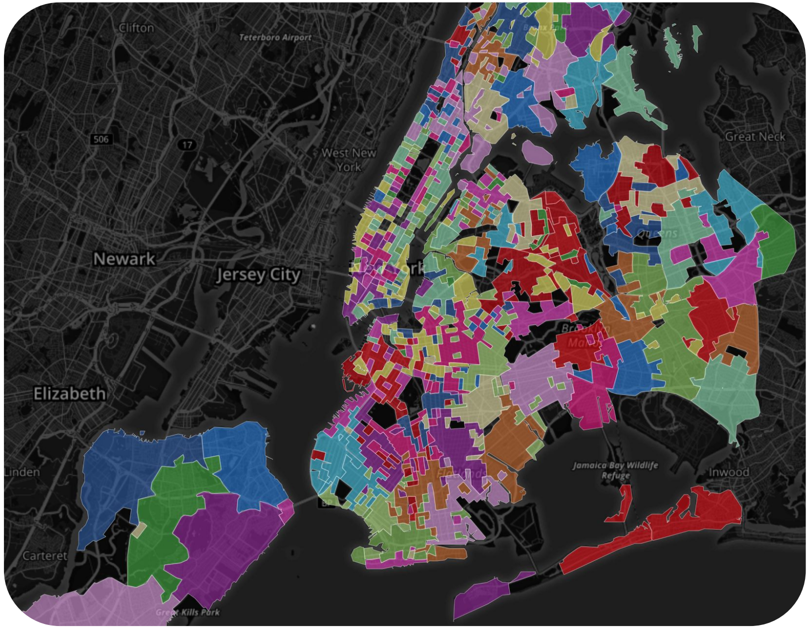
Source: Kings County Census Viewer

TWO SIGMA

data clinic

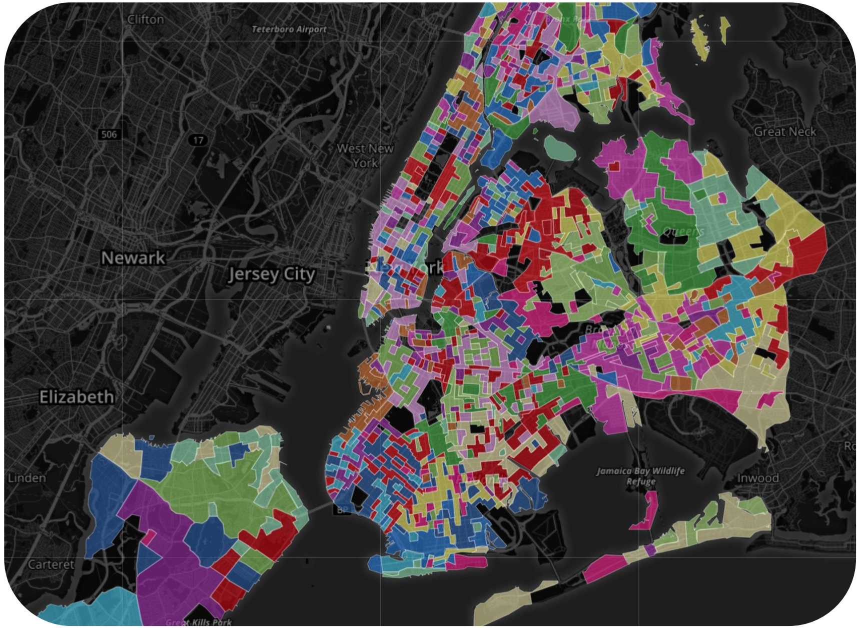
Traditional clustering methods fail

k-means clustering



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

k-means clustering



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

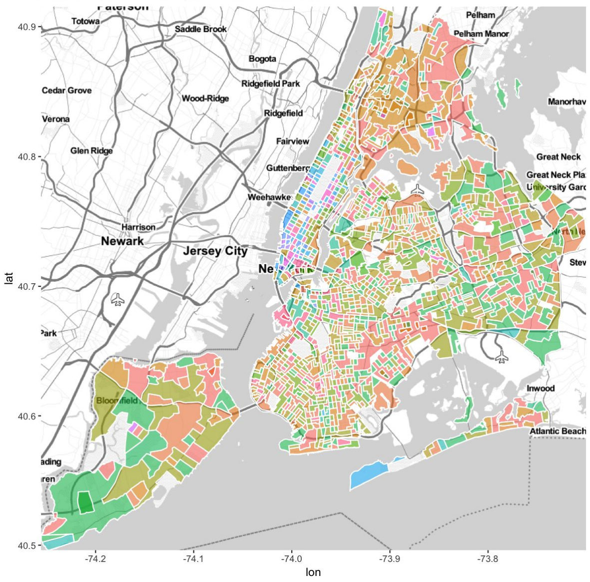
TWO SIGMA

data clinic

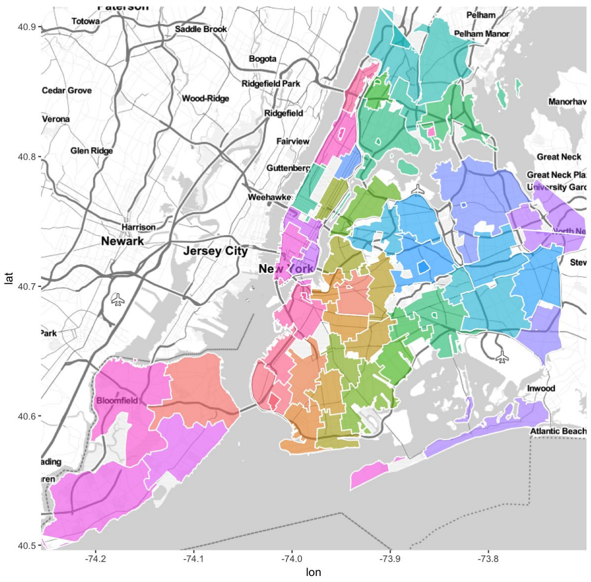
ClustGeo in action

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

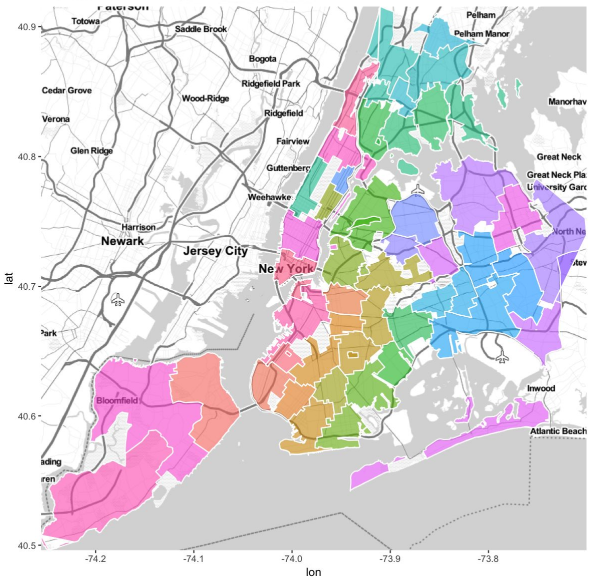
alpha= 0, k = 50



alpha= 0.15, k = 50



alpha= 1, k = 50

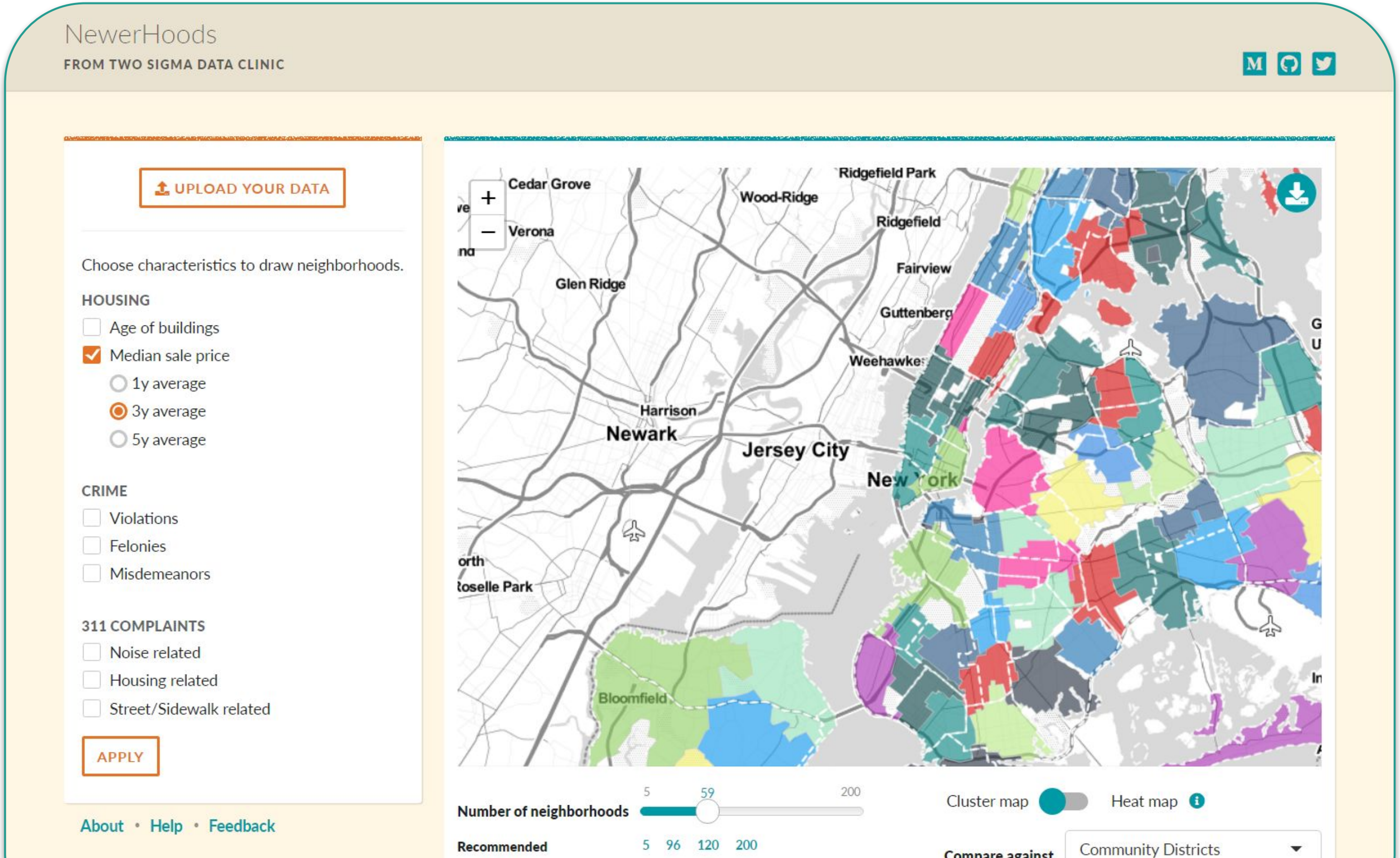


Feature space: mean & sd of price/sq. ft

Geographic space: tract contiguity

TWO SIGMA

data clinic



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

@tsdataclinic

dataclinic@twosigma.com

dataclinic.twosigma.com

thank you

Erin Stein

dataclinic.twosigma.com

@tsdataclinic

dataclinic@twosigma.com