Data Science for Public Policy & Social Impact

Amy Smith | amy.mapsmith@gmail.com

About Me

- Bachelors & Masters in Geography
- Based in San Francisco
- Working with geospatial data for 10+ years
- Most recently a data scientist on Uber's Policy Research & Economics team

What makes Public Policy Data Science unique?

Public Policy Data Science applies data science skills normally used for narrowly focused scientific and business questions on a broad range of social issues that impact the daily lives of community members.

- How do policies impact different groups?
- What are the outcomes of different policies, and do they measure up to goals?
- How feasible and expensive are policies to implement?
- What data-driven insights are needed for decision making?

Who does Public Policy Data Science?

Data Science to inform Public Policy is conducted across sectors and industries

- Academics
- Public Agencies
- Private Companies
- Journalists
- Foundations
- Citizens

At the MIT Policy Hackathon, interdisciplinary teams worked together to find policy solutions to real societal challenges.

Photo: Roxanne Rahnama

Using data science to improve public policy

MIT Policy Hackathon, run by students within MIT's Institute for Data, Systems, and Society, seeks interdisciplinary solutions to societal challenges.

Carnegie Mellon University

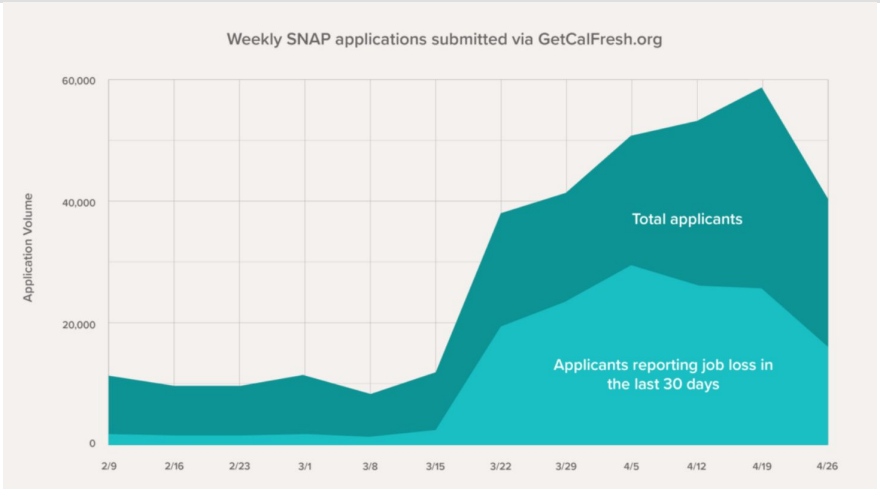
ML Machine Learning Department

Heinz College
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

Are you passionate about social impact and using your AI/ML/Data Science skills to help governments and non-profits be more effective and equitable in improving society?

We are building a new team at Carnegie Mellon University across the Machine Learning Department and the Heinz College of Information Systems and Public Policy focused on using Artificial Intelligence, Machine Learning, and Data Science (and other buzzwords) to have a positive impact on society.

COVID-19 and Food Assistance by the Numbers



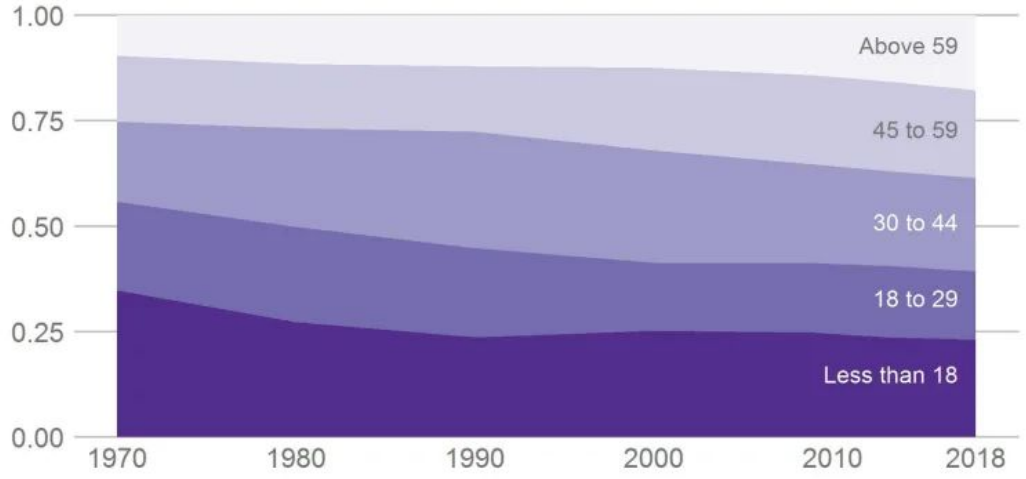
REPORT

As the capital region's population ages, public policies need to adjust

Jaclene Begley, **Leah Brooks**, **Brian J. McCabe**, **Jenny Schuetz**, and **Stan Veuger** · Wednesday, May 20, 2020

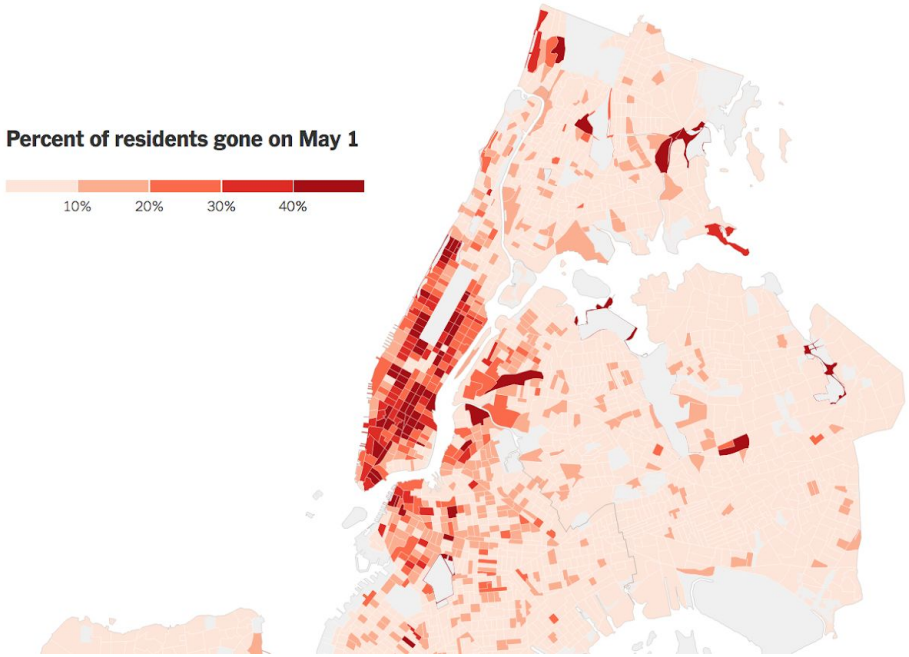
The capital region is aging

Distribution of population by age, 1970-2018



The Richest Neighborhoods Emptied Out Most as Coronavirus Hit New York City

By Kevin Quealy May 15, 2020

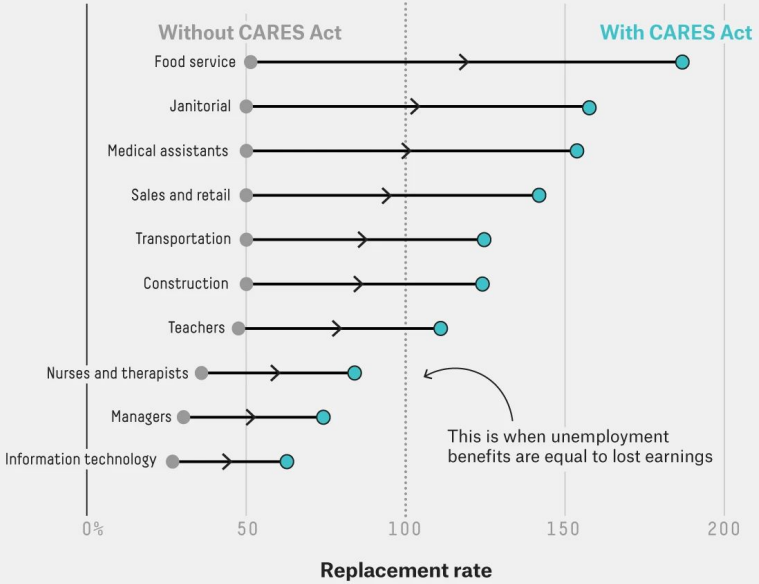


MAY 15, 2020, AT 6:00 AM

Many Americans Are Getting More Money From Unemployment Than They Were From Their Jobs

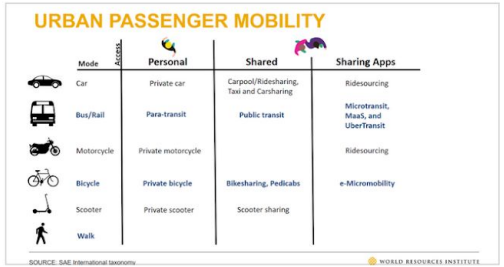
Some workers are making more on unemployment

Estimated percentage of earnings replaced by unemployment benefits for the median unemployed worker in 10 common occupational fields



Under the Hood

Insights and updates from the Uber public policy team



Measuring Mobility for Carbon Efficiency

Carbon intensity metrics for the transportation sector

Michiko Namazu

Sep 26, 2019 · 7 min read

Sharing the Road — Travel Efficiency

Successful cities are crowded places. Keeping them successful means making better use of their limited physical space by managing travel...

Jonathan Wang

Sep 26, 2019 · 8 min read

A Step Forward On Sustainability

Read below to learn about our new 100% renewable electricity goal, new transit and EV efforts, commitment to report on the impact of rides...

Adam Gromis

Sep 26, 2019 · 6 min read

Who tools do Data Scientists in Public Policy use?

The Usual Suspects

- Python
- R
- SQL
- PostGIS
- QGIS
- Curiosity
- Common Sense

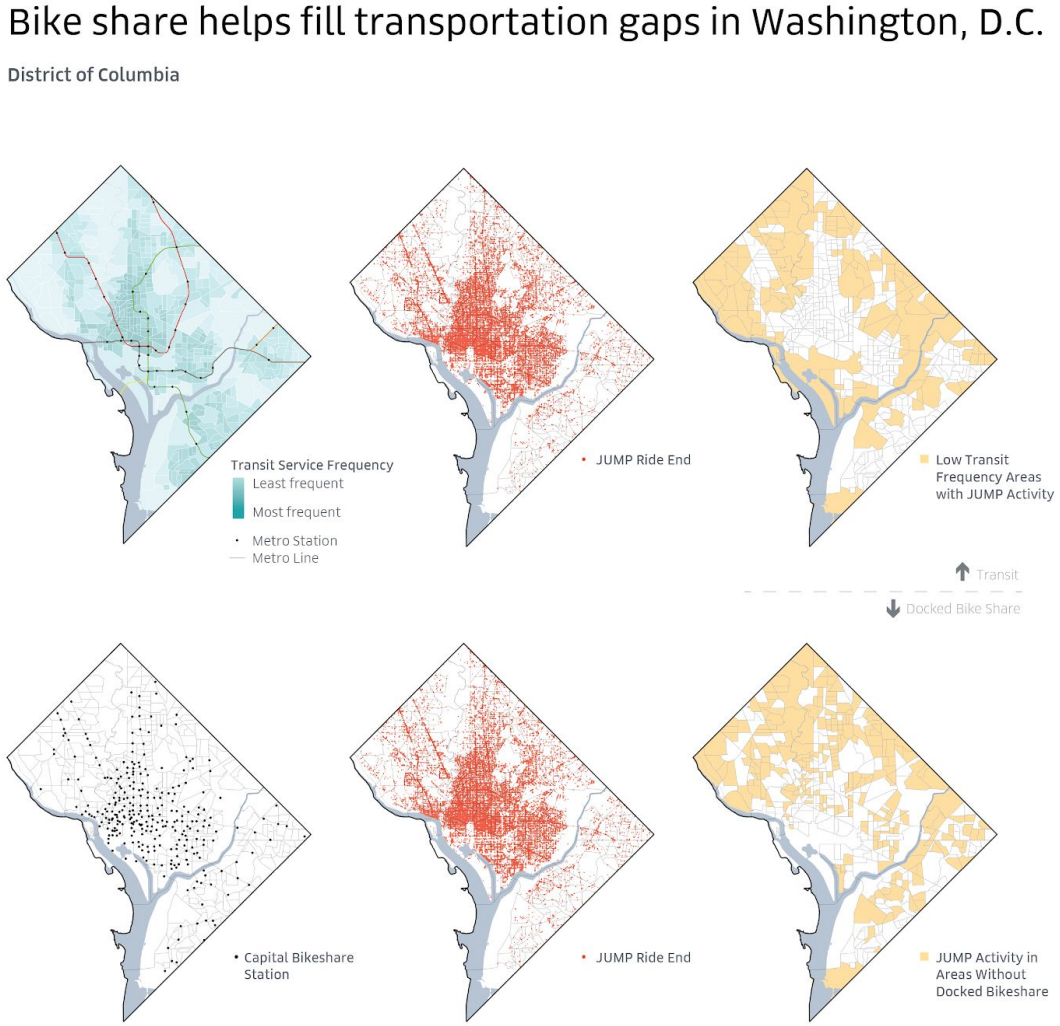
Why “Spatial” Matters

Equitable Bike Share Across Cities

Uber Under the Hood Follow
Mar 28, 2019 · 4 min read

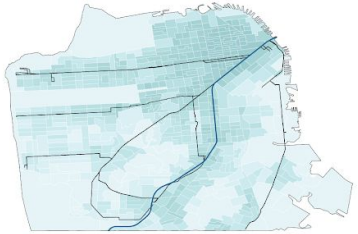
By Malcom Glenn, Head of Global Policy, Accessibility and Underserved Communities, and Amy Smith, Policy Research Data Scientist

When JUMP joined the Uber team last year, part of the reason we were so excited was that we believed that dockless e-bikes could improve equitable access to low-cost transportation throughout cities, particularly parts of cities that were traditionally poorly-served by existing transportation options.



Bike share helps fill transportation gaps in San Francisco

San Francisco, California



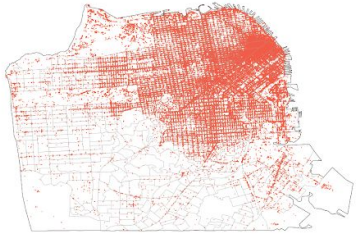
Transit Service Frequency

Least frequent

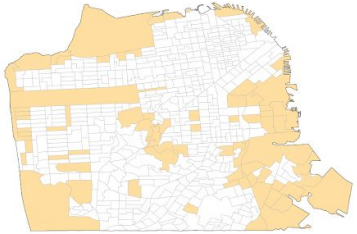
Most frequent

— Bay Area Rapid Transit

— SF Muni Metro and 38 Geary Bus Lines



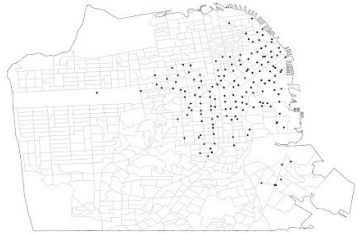
• JUMP Ride End



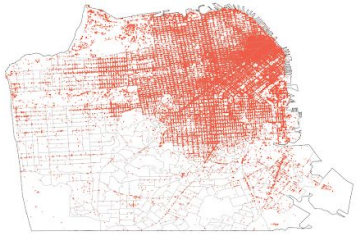
■ Low Transit Frequency Areas with JUMP Activity

↑ Transit

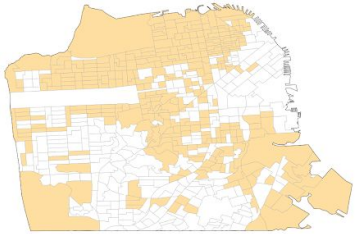
↓ Docked Bike Share



• Ford GoBike Station



• JUMP Ride End

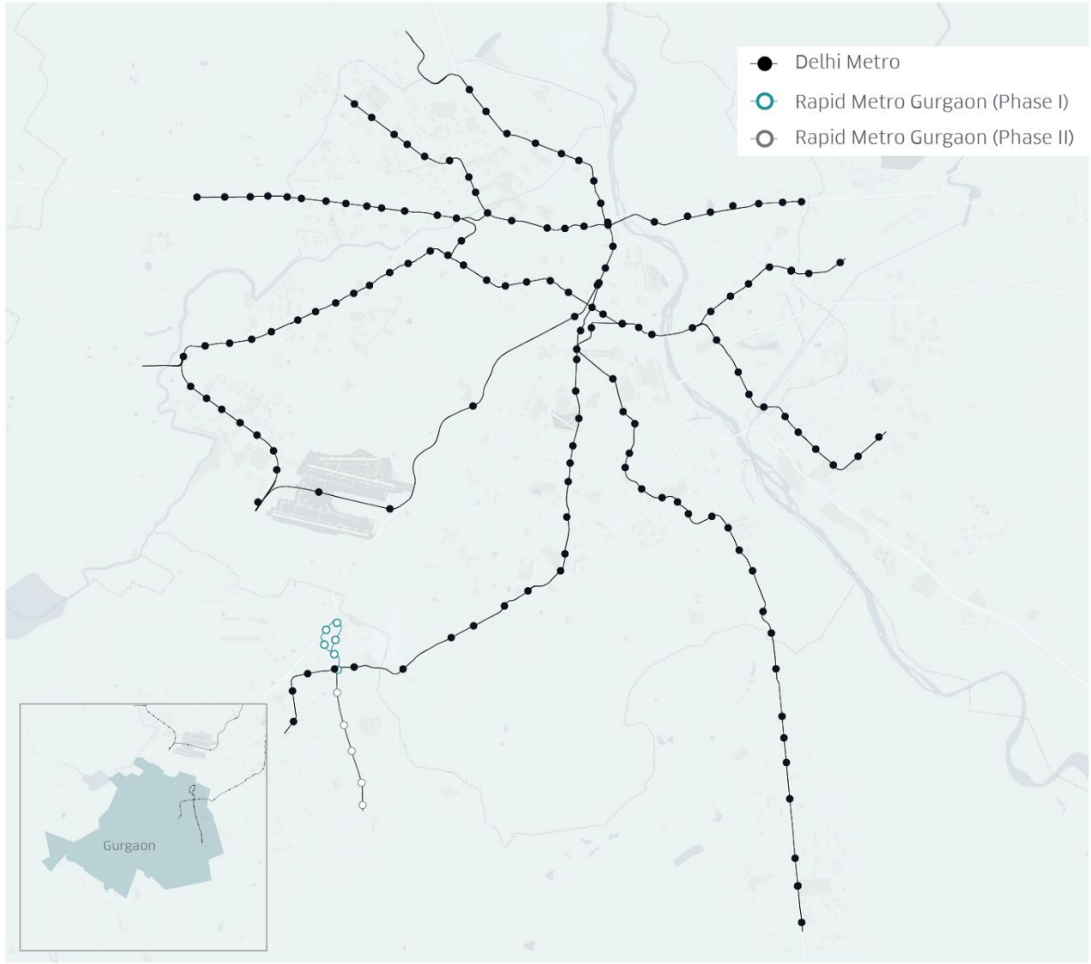


■ JUMP Activity in Areas Without Docked Bikeshare

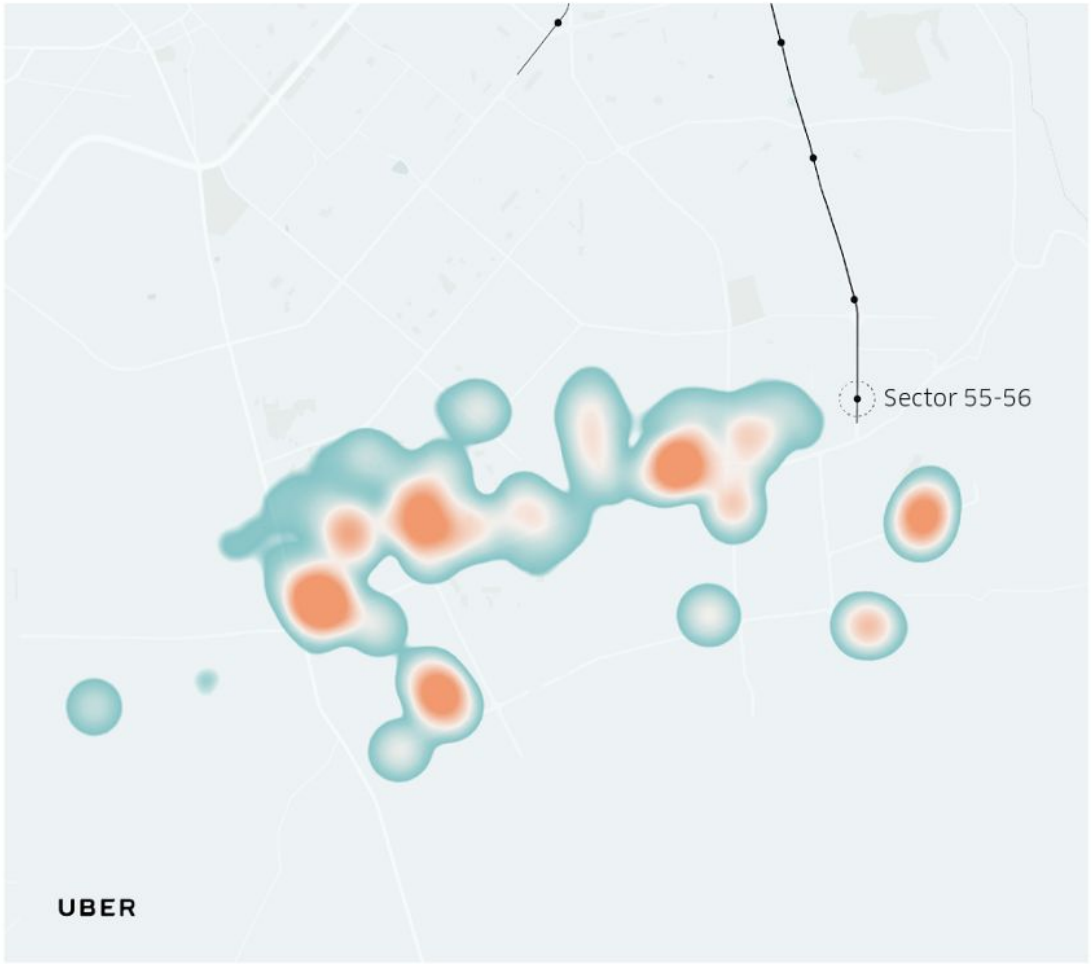
Extending Public Transport in India's Millennium City

Amy Smith Follow
Dec 20, 2017 · 3 min read

From the front entrance of Sikanderpur Metro Station, one can spot travelers from all directions approaching the station. Some are on foot, while others prefer to make the journey on autorickshaw, two-wheeler, minibus, taxi or Uber. The station is a nerve center of multimodal transportation in Gurugram, the financial and industrial hub just south of New Delhi — and for travelers coming from India's capital city on the Delhi Metro, Sikanderpur Station also serves as the transfer point between the Metro and Gurugram's intra-city elevated light rail system.



Basemap courtesy of OSM and Mapbox.



Uber extends the reach of Gurgaon Rapid Metro.

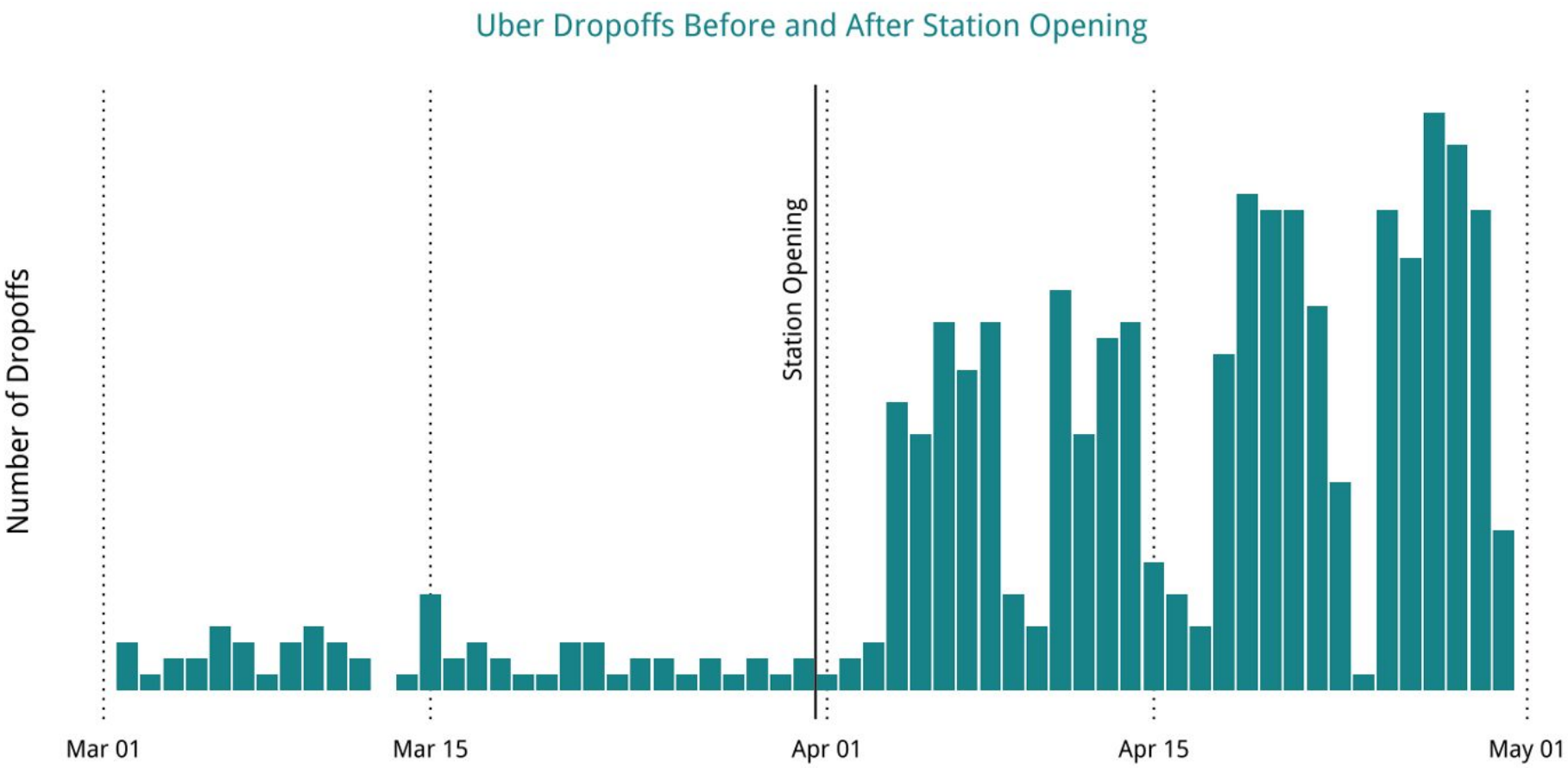
Gurgaon, Haryana

The color on the map indicates the magnitude of Uber pickups for trips that dropped off near the Sector 55-56 Rapid Metro station.

Pickups

Few — Many

Data from April 1st to April 30th. Completed trips only. Basemap courtesy of OSM and Mapbox.


Uber Dropoffs Before and After Station Opening
Number of Dropoffs
Station Opening
Mar 01
Mar 15
Apr 01
Apr 15
May 01

Getting Started

Ask Questions

What do I care about?
What questions have been asked, and where are there gaps?
What's *my* question?
Where can I find the data?

Collect Data

Some Examples
Open Data Portals
Census and American Community Survey
EPA Smart Location Database
Bureau of Labor Statistics
NASA (satellite imagery)
General Transit Feed Specification
OpenStreetMap
Natural Earth Data
World Bank
Bureau of Transportation Statistics
National Household Travel Survey

I didn’t find any data, is it hopeless?

❌ **Yes**

>>>

I guess we’ll never know

✅ **No**

>>>

What are the important questions I should be asking next, why are there gaps, what are the implications of not having good answers to this question, who wants to solve this problem with me?

The data didn’t tell me what I thought it would, should I toss the whole analysis?

❌ **Yes**

>>>

No one will publish this anyway...

✅ **No**

>>>

Lack of significant findings are useful – save someone a few steps in their research process by sharing your underwhelming results. If the results will make someone unhappy, that’s OK too.

Demo

I care about walkable, livable streets and easy, equitable access to resources in my city. What data can I use to identify areas that could benefit from improvement?

Economy and Community ×

Show All...

View Types

Calendars

Charts

Data Lens pages

Datasets

External Datasets

Registered Business Locations - San Francisco

Economy and Community

Dataset

This dataset includes the locations of businesses that pay taxes to the City and County of San Francisco. Each registered business may have multiple locations and each location is a single row. The Treasurer & Tax Collector's

More

Tags business

API Docs

Updated
May 23, 2020

Views
226,097

Active Business Locations

Economy and Community

Filtered View

Registered business locations in San Francisco maintained by the Office of Treasurer-Tax Collector, including business locations that have been sold, closed, or moved out of San Francisco.

Less

Tags business

API Docs

Updated
May 23, 2020

Views
10,761

Registered Business Locations - San Francisco

Economy And Community

View Data | Visualize | Export | API | ...

This dataset includes the locations of businesses that pay taxes to the City and County of San Francisco. Each registered business may have multiple locations and each location is a single row. The Treasurer & Tax Collector's Office collects this data through business registration applications, account update/closure forms, and taxpayer filings. The data is collected to help enforce the Business and Tax Regulations Code including, but not limited to: Article 6, Article 12, Article 12-A, and Article 12-A-1. http://sftreasurer.org/registration

Less

Featured Content Using this Data

Public Map of Registered Businesses - San Francisco

May 23, 2020 1,958 Views

Download Registered Business Locations - San Francisco

Download Registered Business Locations - San Francisco for offline use in other applications.

CSV | KML | Shapefile

Additional Formats

CSV for Excel | KMZ | TSV for Excel

CSV for Excel (Europe) | RDF | XML

GEOJSON | RSS

Import the data into a PostgreSQL database and join the active businesses to the registered businesses

```
CREATE TABLE sf_businesses AS (
    SELECT DISTINCT r.dba_name AS bus_name
            , r.certificat AS bus_id
            , r.lic AS lic_code
            , r.lic_code_d AS lic_name
            , r.naic_code AS naic_code
            , r.naic_code_ AS naic_name
            , CASE WHEN "Business Account Number" IS NOT NULL
                    THEN 1 ELSE 0 END AS active
            , ST_Transform(geom, 32611) AS geom
    FROM sf_registered_businesses r
    LEFT JOIN sf_active_businesses a
            ON a."Business Account Number" = r.certificat
);
```

Transform the geometry so you can apply clustering distances in meters (vs. decimal degrees)

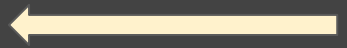
Create a new subset of only food-related businesses that someone would be able to walk to and purchase food.

```
CREATE TABLE sf_food AS (
    SELECT DISTINCT bus_name
            , bus_id
            , lic_code
            , lic_name
            , naic_code
            , naic_name
            , active
            , geom
    FROM sf_businesses
    WHERE 1=1
            AND active = 1
            AND (
                    LOWER(lic_name) LIKE '%food%'
                    OR LOWER(lic_name) LIKE '%produce stand%'
                    OR LOWER(lic_name) LIKE '%certified farmers markets%'
                    OR LOWER(lic_name) LIKE '%restaurant%'
                    OR LOWER(lic_name) LIKE '%take-out establishment%'
                    OR LOWER(lic_name) LIKE '%supermarkets%'
            )
            AND LOWER(lic_name) NOT LIKE '%mobile food prep unit%'
            AND LOWER(lic_name) NOT LIKE '%caterer retail food vehicles%'
            AND LOWER(lic_name) NOT LIKE '%school%'
);
```

Filter the data using the documentation provided with the open dataset.

Use a clustering algorithm to identify spatial clusters of food resources. Try out different thresholds to see what works.

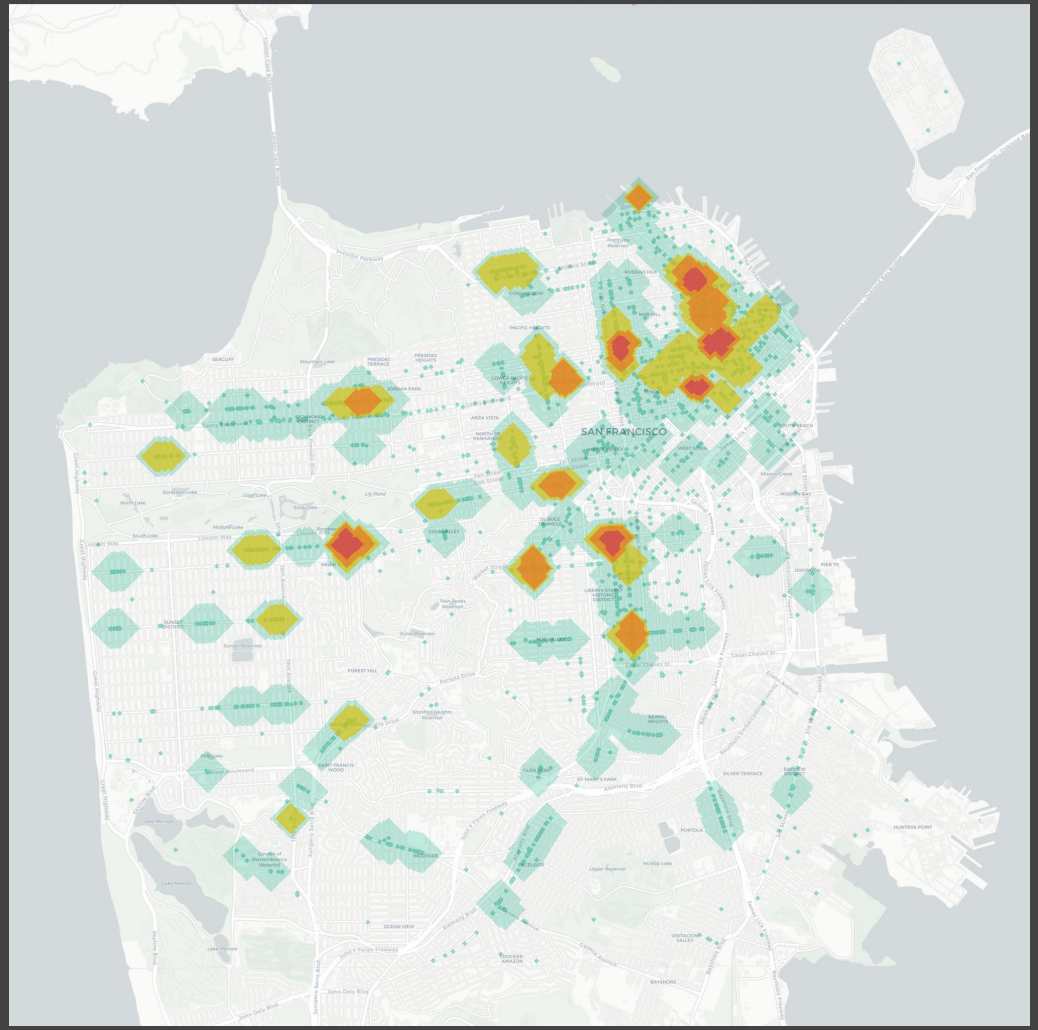
```
DROP TABLE clustered;
CREATE TABLE clustered AS (
    SELECT bus_name
    , bus_id
    , lic_code
    , lic_name
    , naic_code
    , naic_name
    , active
    , ST_ClusterDBSCAN(geom, eps := 100, minpoints := 5)
            over () AS cid, geom
           FROM sf_food
    WHERE 1=1
);
```



ST_ClusterDBSCAN will cluster geometries based on a search radius and minimum number of points.

Map it Out

Food Resource Clusters in San Francisco



Food Resource

Food Cluster

High Density

Low Density

Ask More Questions

- Are there sufficient pedestrian facilities in areas with high density food clusters?
- Who’s within a walkable distance in these areas, where are there gaps in easy access to food resources by foot or other modes of transportation?
- What additional data do I need to paint a full picture?

THANK YOU!

amy.mapsmith@gmail.com