MEASURING ACCOUNTABILITY

A QUANTITATIVE ANALYSIS OF DATA TRANSPARENCY IN LOCAL GOVERNMENTS

1 QUESTION

The Obama administration brought attention to Open Data in the United States by creating the Memorandum on Transparency and Open Government, which states, **Information maintained by the Federal Government is a national asset. My Administration will take appropriate action, consistent with law and policy, to disclose information rapidly in forms that the public can readily find and use.** The Digital Accountability and Transparency Act, signed in 2014, aims to increase access to information on federal expenditures and has led to the creation of nationwide open data platforms such as data.gov. While this is a major step forward, the law does not mandate such transparency at the state and local level. Since open data is vital in evaluating the effectiveness of public officials and holding them accountable, we wanted to analyze how "open" data at the local level is.

By Aura Barrera, Akhil Jalan, Ameet Rahane and Clara de Martel

2 METRIC

To gather data in a standardized manner, we built a web scraping tool. This tool accepts the open data platform of any city that uses Socrata, software which offers standardized data hosting. In order to get each dataset page, we created an algorithm that deterministically goes through each navigation page of the open data platform, checks if the link fulfills the trait of a dataset and creates a list of all of those links.

To extract most other information, we applied regular expressions (RegEx) on the source code of the necessary webpage. For any remaining information, we created a script that would extract information from the 'meta' section of the dataset file. We used a Python JSON reader to import the data files, which we then used to extract relevant information about the view count, recency, etc...

1 QUANTITY

We calculate the number of datasets per city by parsing the HTML of the data homepage and scale by government size, to normalize and compare.

2 RECENCY

This metric takes into account the date last updated (in days) and the date of the "present". We scale last update time according to their recency so that the oldest datasets have a smaller effect on the score.

Quantity	Ranking based on number of datasets published, scaled according to size of city government
Recency	Ranking based on average amount of time passed since updating of datasets on the platform
Accessibility	Measured using average dataset download count and number of views
Breadth	Calculation of the spread and relevancy of datasets available using tags and dataset titles
Quality	Measure of dataset machine readability and adherence to data delivery standards

3 ACCESSIBILITY

From information published in each of the JSON data files, we extract number of views and downloads and over the datasets.

4 BREADTH

Using a Natural Language Processing (NLP) benchmark known as GloVe similarity,[1 2] we modeled our data categories on NYC open data.[3] We extracted tags from the city government websites and computed similarity scores (0 - 1) — by averaging inner products of the embedded category word-vectors and the website tag word-vectors — for each of the predefined categories.

5 QUALITY

We use the Project Open Data Dashboard's open-source schema matching validator on several randomly selected datasets from each city to standardize a metric for data quality.[4]

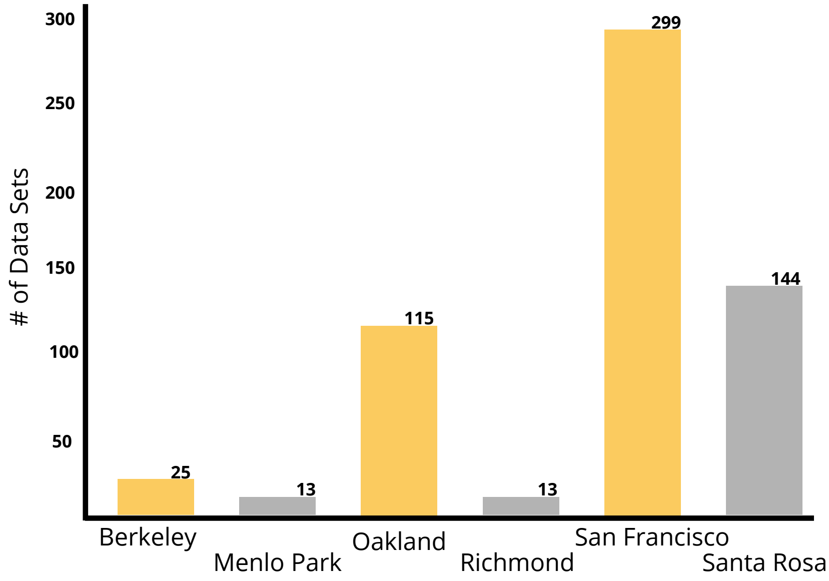
3 RESULTS

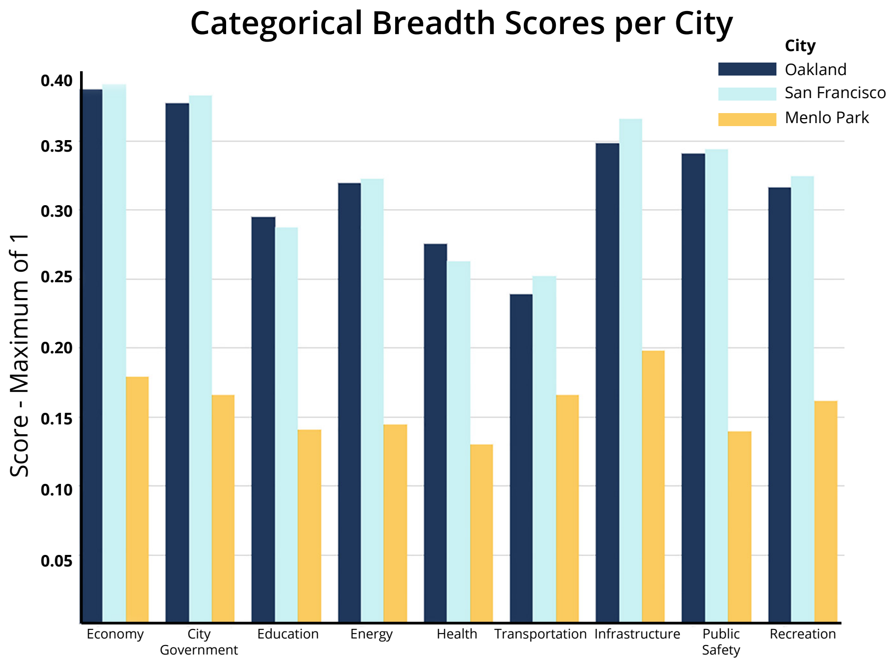
We focus our study on the Socrata data platforms of local governments in the Bay Area. We compare the results from these cities to that of randomly selected cities, who also use Socrata, across the United States.

Disclaimer: Not all data is shown here, further information published on our website and will be available on April 30th.

Santa Rosa
Richmond
Berkeley
Oakland
San Francisco
Menlo Park

Quantity of Data Sets per City





On our website you will be able to find:

◊ Downloadable CSV for all cities containing scores for each part of the metric

◊ Downloadable CSV for each city with metadata on each dataset

◊ Data visualizations

◊ Comprehensive write up about our methodology

Here we include two graphs demonstrating some of the data collected for Quantity and Breadth.

4 CONSIDERATIONS

Insight into what our metrics are missing, how they can be improved, and what lessons we learned along the way:

RECENCY

ISSUES

To address the timeline of the data update history, it would be more accurate to measure the frequency of update rather than just the last one. Additionally, we would want to identify time insensitive datasets, so that we do not include them in this score.

SOLUTIONS

We will need to scale our results by an update frequency ratio. We could implement an algorithm that reads datasets by title and clusters them into categories that indicate whether they need to be updated regularly or not.

QUALITY

ISSUES

This method only checks validity and machine readability of the data. Both of these factors are important in their own right, but it is in no way representative of data quality. Other factors, as mentioned in the DAMA UK Working Group on “Data Quality Dimensions should measure completeness, accuracy, and consistency.[5]

SOLUTIONS

Completeness and accuracy seem hard to validate without some other source for the data. While we may be able to find other sources for some datasets, it would be quite a technological feat to do this automatically for all datasets. Thus, it would be better to validate datasets working with the city governments in order to traceably solve this problem. To measure for consistency we can create a regression of amount of data over time and formulate a

ACCESSIBILITY

ISSUES

Other potential factors we should address have to do with the reachability from a search engine, reachability within the local city website itself, and user-friendliness.

SOLUTIONS

For reachability from a search engine and within the local city website, we could generate a URL graph that, given a few search terms, goes through all of the links. The graph would stop when it reaches the open data protocol page

PRIVACY

To address potential privacy risks and social harms, we followed a framework from Seltzer and Anderson regarding considerations that can prevent potential misuse of the data. Some factors to consider are populations’ power, data’s relation to stigma, whether sampling is used, level of coercion, procedural legitimacy, climate/context—political, cultural, and economic.[6] We were careful about the legitimacy of our results since we are not experts in the subject area.

CITATIONS

1. GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/projects/glov
2. Pennington, Jeffrey, Richard Socher, and Christopher Manning. “Glove: Global vectors for word representation.” Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014.
3. New York City Open Data platform, https://opendata.cityofnewyork.us/
4. Project Open Data dashboard, https://labs.data.gov/dashboard/validate
5. The Six Primary Dimensions for Data Quality Assessment, https://www.whitepapers.em360tech.com/wp-content/files_mf/1407250286DAMAUKDQDimensionsWhitePaperR37.pdf.
6. Seltzer, W., & Anderson, M. (2001). The dark side of numbers: The role of population data systems in human rights abuses