Open Data Had Better Be Data-Driven

By Thomas Levine

Dada Artist

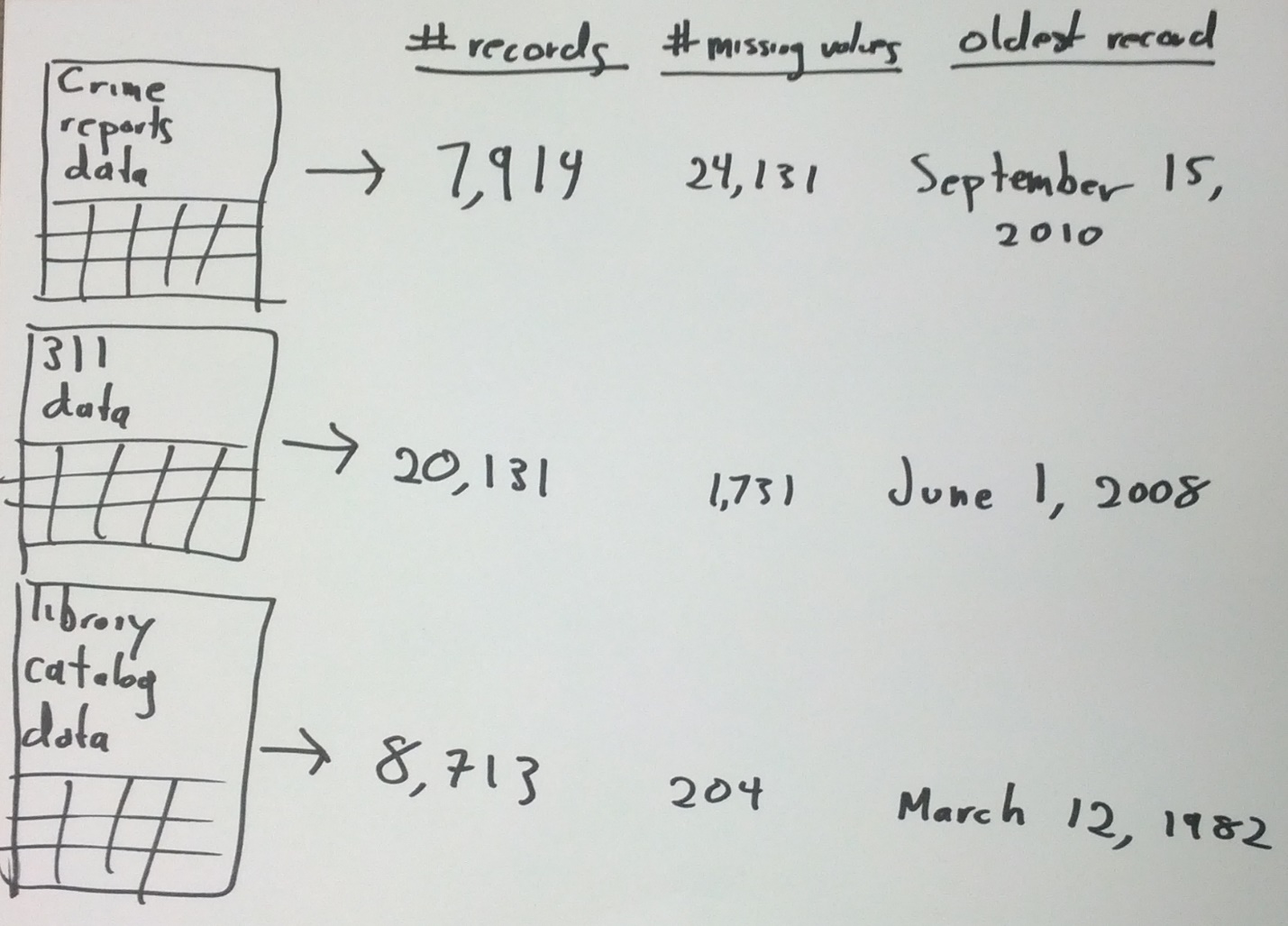
We've been opening government data for some time now. Without realizing it, we've amassed some rather rich data about people publish and consume open data. With these data about the use of open data, it’s possible to use data-backed benchmarks, projections, and decisions in our open data strategies strategies, and this can make our approaches to open data more systematic, logical, and obtainable.

Curious about what the data behind open data can teach us? Read on.

**Data, Data, Everywhere**

We have data about open data, but it always starts out in formats that are not convenient for this sort of study of open data. So we first need to turn this raw data into a dataset. In my mind, a dataset is a collection of things, with some consistent properties describing each of the things. (We often represent datasets as [tables](http://www.datakind.org/blog/whats-in-a-table/).) We're going to treat each dataset as a thing inside our collection of many datasets.

Every dataset has basic, easy-to-extract properties like the number of records it contains, and the size it takes up on a hard drive. We can come up with more complicated too, like the number of missing values, and the date of the oldest record.



[Caption: By taking basic information from individual datasets, you get data that you can use to create a new dataset.]

In addition, when a dataset is published online, it creates metadata – web page information about the dataset and its use. Metadata provides details like when it was first published, who uploaded it, and how many times people have downloaded it since then.

"name" : "Open Meetings Calendar","averageRating" : 0,

"category" : "Government Administration",

"createdAt" : 1337110426,

"description" : "Schedule of Missouri Open Meetings",

"displayType" : "calendar",

"downloadCount" : 68,

"numberOfComments" : 0,

"publicationDate" : 1384531257,

"publicationStage" : "published",

"rowsUpdatedAt" : 1384531248,

"rowsUpdatedBy" : "8xqn-4t42",

[Caption: [Metadata](https://data.mo.gov/views/mahp-izvx.json) about Missouri's [open meetings calendar](https://data.mo.gov/Government-Administration/Open-Meetings-Calendar/mahp-izvx).]

By collecting some of these simple properties and metadata from each dataset, it’s possible to create a dataset about the publication and use of other datasets – what I call a super-dataset.

**Datasets as Data Points**

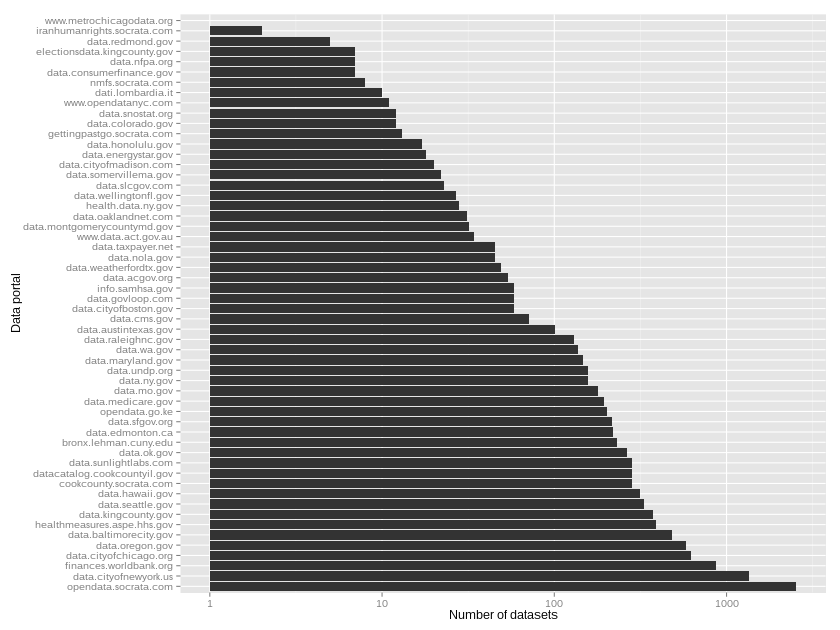
A super-dataset compiles information about each dataset – when it was published, what kind of information it contains, etc. – into a single row, creating a record (i.e. a data point) about that dataset.

With this setup, you can perform all kinds of functions and analyses about the publication and use of datasets: you can look at [how many datasets are on different catalogs](http://thomaslevine.com/!/socrata-deduplicate/), how data is [queried and reported](http://www.chriswhong.com/nycopendata/), what [licenses](http://thomaslevine.com/!/open-data-licensing) datasets have, and how many there are of a [certain category](http://thomaslevine.com/!/missouri-data-licensing/). Here are some examples of analyses I have done.

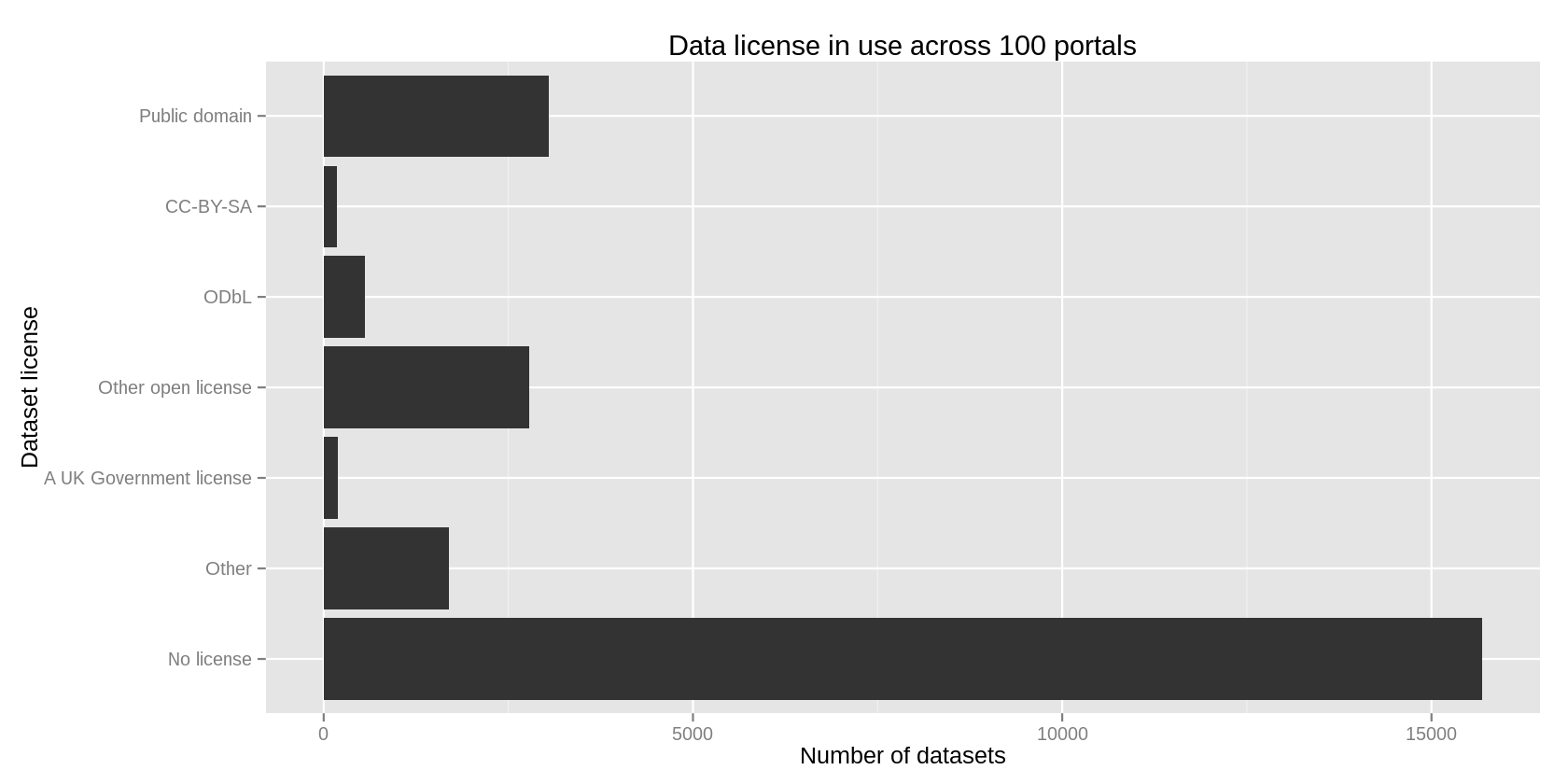
**Number of Datasets:** I compared the number of datasets on some various government data portals that run Socrata's open data catalog software. I found that New York City, Chicago, and the state of Oregon are the governments with the most datasets.

**Licensing:** I also used my super-dataset of properties and metadata to look into this question: What licenses do people apply to their open data? I discovered that many portals favor public domain or some form of open license, but most list no license at all.

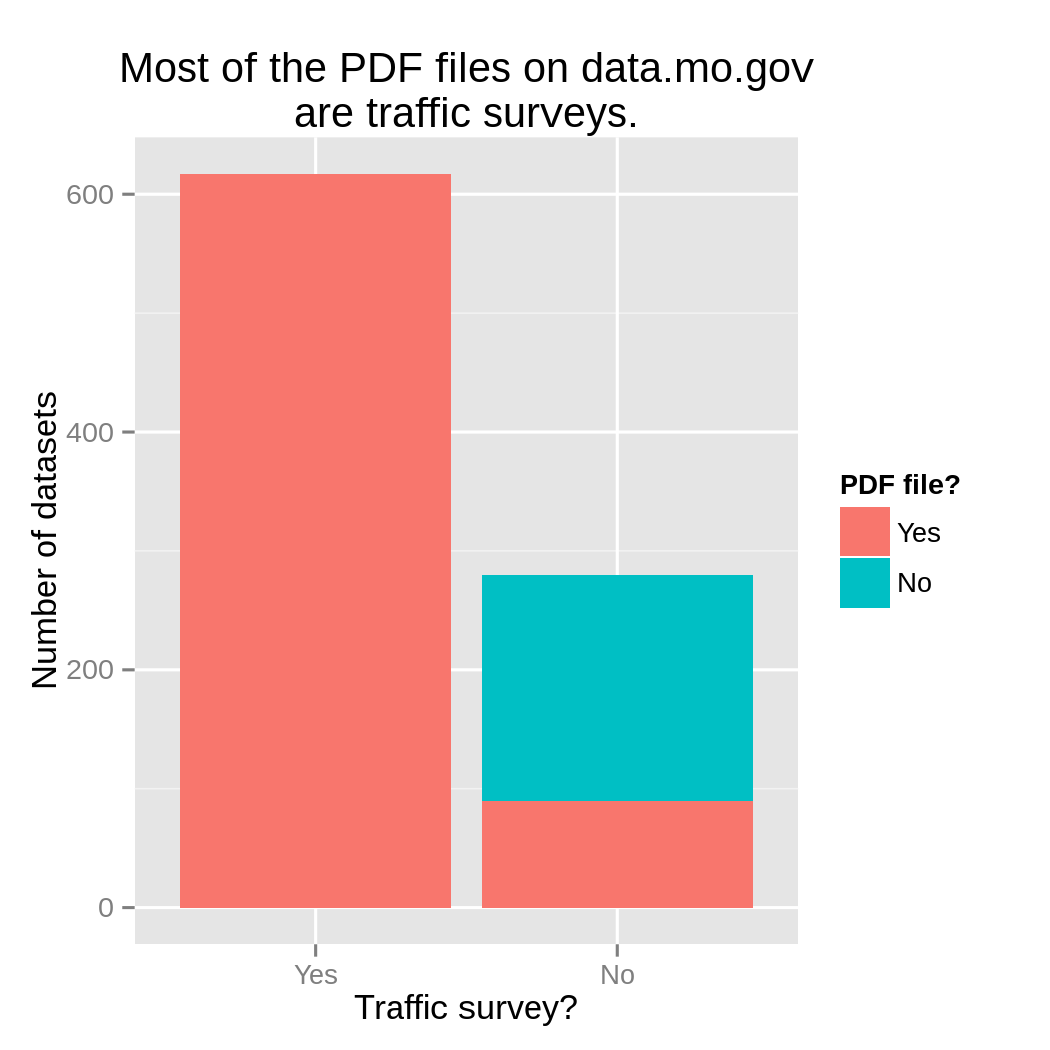
**Groupings of datasets:** By looking at similarities in the titles, schemas, and other metadata of datasets, I determined what sorts of data were putting on their open data portals and how different datasets were related to each other.



[Caption: This super-dataset graphs the number of datasets on various government data portals]



[Caption: This graph, fueled by data from a super-dataset, reveals many portals favor public domain or some form of open license, but most list no license at all.]



[Caption: By filtering datasets by file type, category, etc., and comparing them to metadata records such as views, downloads, comments, you can determine which types of records are of biggest interest to the public.

**Metadata as Data**

Even though the word "metadata" contains the word "data," people don’t typically think of metadata as something to analyze.

Metadata is often invisible. If we can see it, it might just look like background information about web pages and their contents. For example, Socrata’s open data platform displays metadata about each dataset, such as community rating, number of visits, number of downloads, etc.

When you use metadata to populate records in a super-dataset, that metadata becomes data that we can see, analyze, and learn from.



[Caption: Metadata from Socrata’s platform converted into data in a super-dataset.]

**Why This Matters**

There are numerous case studies, by organizations such as [Code for America](http://beyondtransparency.org/), [CKAN](http://ckan.org/case-studies/), [Open Data Institute](http://theodi.org/case-studies), about how to open up government data, but these are based strongly on personal experiences, not precise, quantitative statistics. There is [comparatively little](http://thomaslevine.com/open-data) work that uses data to produce guidelines or best practices for the opening of data.

Open data experts often talk about making use of open data and building products from it, but a super-dataset accomplishes something different. It reveals information about how people publish and use open data. It paints the bigger picture of what's going on, what's working, and what isn't.

Imagine what we can do with this data-backed understanding of our open data! We can find out what has been done before and what has worked, allowing open data publishers plan their releases more strategically. We can measure release strategies against solid, quantitative statistics to make sure that they are achieving our goals. We can even use these findings to build products that people interact with open data. These are just some of the possibilities that emerge when we think of data catalogs as a datasets of datasets.