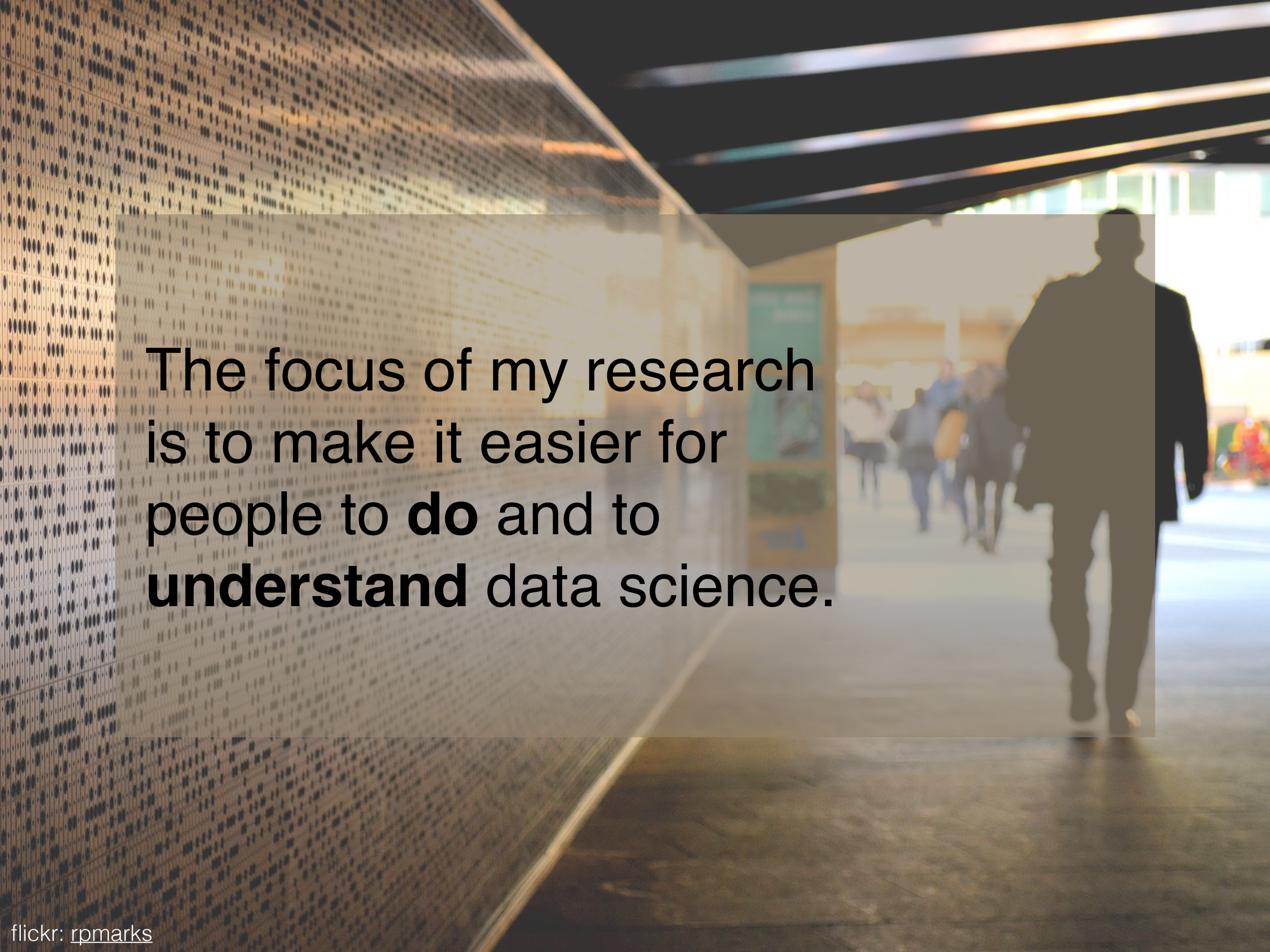
Making spatial aggregation more transparent

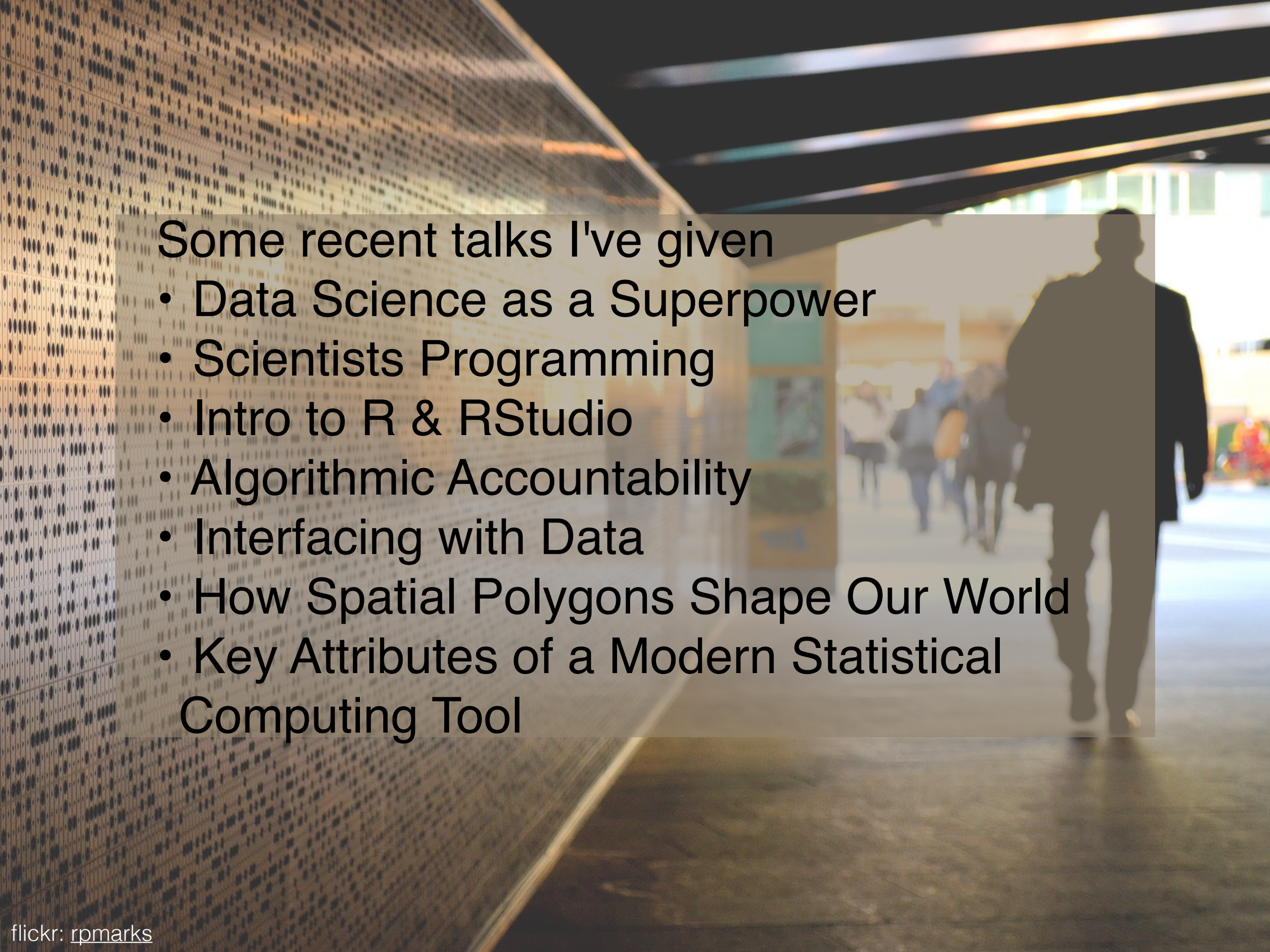
Amelia McNamara @AmeliaMN
Currently: Program in Statistical and Data Sciences, Smith College, Northampton MA
Fall 2018: Department of Computer & Information Sciences, University of St Thomas, St Paul, MN

flickr: bering_land_bridge


I have many research interests, including
• statistical computing
• reproducible research
• statistics education
• data visualization
• spatial statistics
• citizen science
• participatory sensing
flickr: rpmarks



flickr: rpmarks



Some recent talks I've given

- Data Science as a Superpower
- Scientists Programming
- Intro to R & RStudio
- Algorithmic Accountability
- Interfacing with Data
- How Spatial Polygons Shape Our World
- Key Attributes of a Modern Statistical Computing Tool

flickr: rpmarks

Gather your data

A histogram is based on a collection of data about a numeric variable. Our first step is to gather some values for that variable. The initial dataset we will consider consists of fuel consumption (in miles per gallon) from a sample of car models available in 1974 (yes, rather out of date). We can visualize the dataset as a pool of items, with each item identified by its value—which in theory lets us "see" all the items, but makes it hard to get the gestalt of the variable. What are some common values? Is there a lot of variation?

Sort into an ordered list

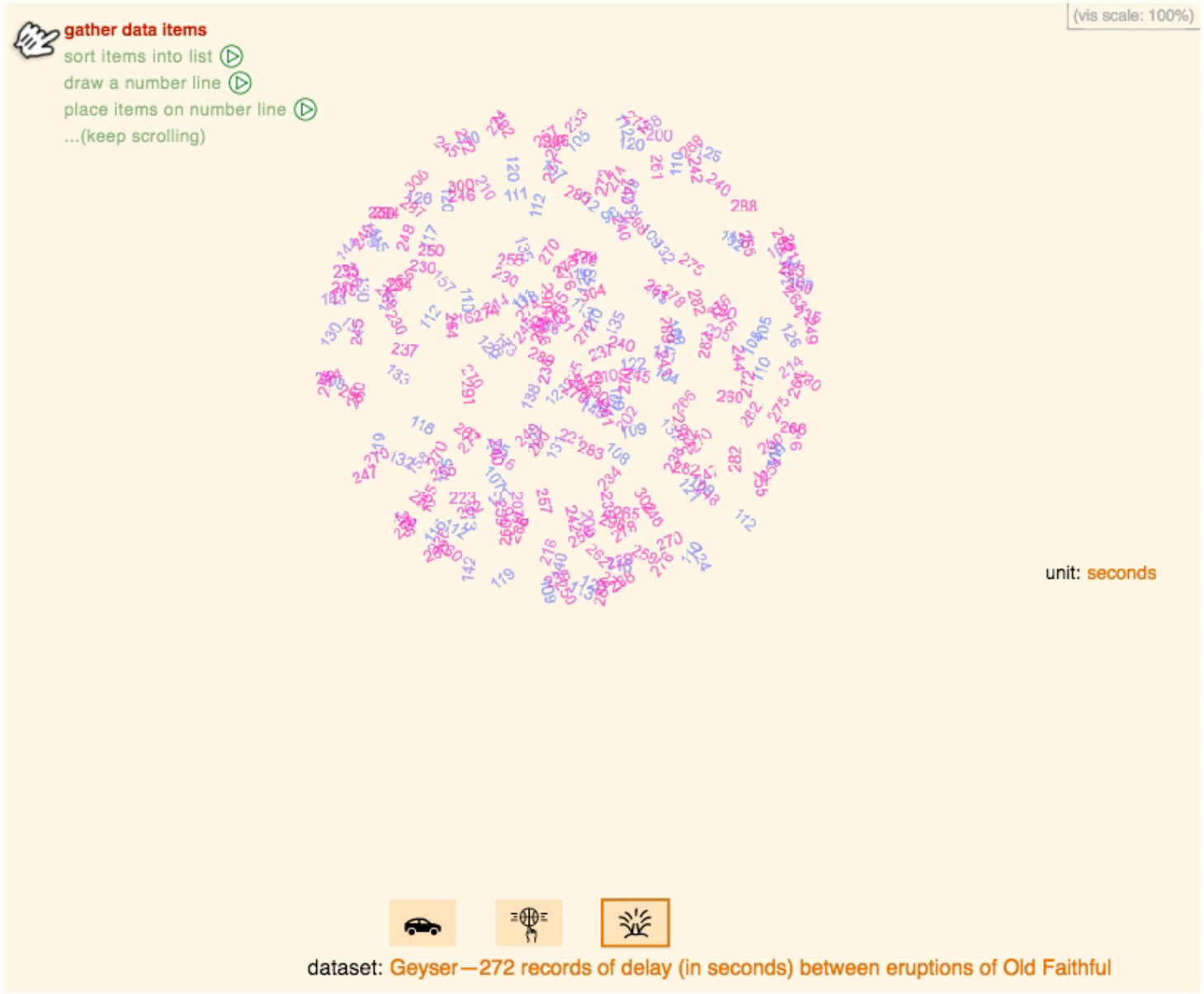
A useful first step towards describing the variable's distribution is to sort the items into a list. Now we can see the maximum value and the minimum value. Beyond that, it is hard to say much about the center, shape, and spread of the distribution. Part of the problem is that the list is completely filled; the space between any two items is the same, no matter how dissimilar their values may be. We need a way to see how the items relate to each other. Are they clustered around a few specific values? Is there one lonely item, with a value far removed from all the others?

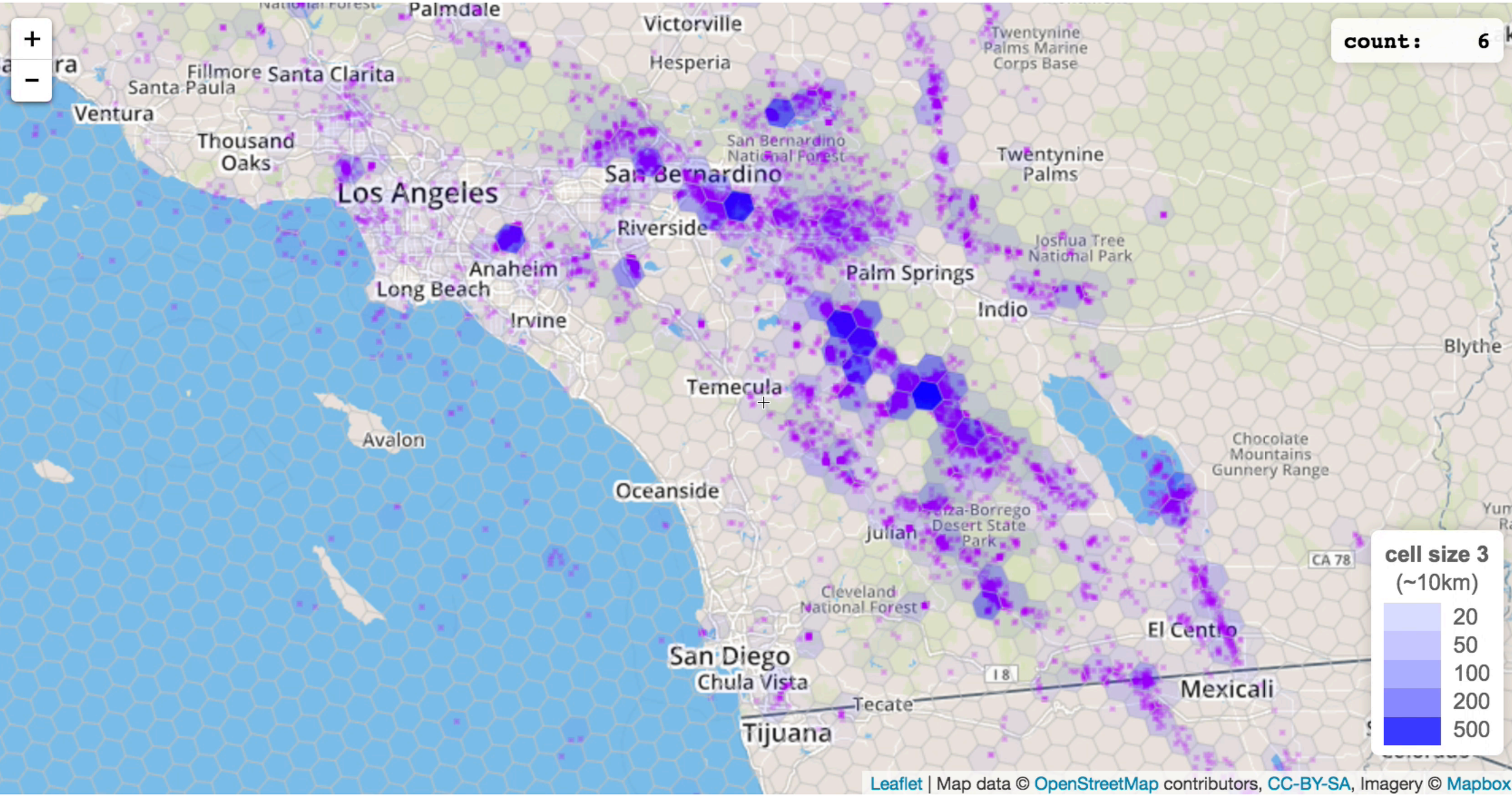
Draw the number line

A common convention is to use a number line, on which higher values are displayed to the right and smaller (or negative) values to the left. We can draw a line representing all possible numbers between the minimum and maximum data values.

Add data to the number line

Now, we map each item to a dot at the appropriate point along the number line. In our visualization we draw the path followed by each item on its way from the list to the line, helping to reveal how adjacent list items end up close or far apart on the number line





Lunzer and McNamara, http://bit.ly/Spatial_Agg

Thank you

flickr: bering_land_bridge