Data Storytelling Studio

analyzing data

CMS.631/831
Rahul Bhargava
Agenda

- [10] basic stats
- [10] grad reading presentation
- [20] a taxonomy of story types
- [20] WordCounter storyfinding activity
- [15] tool demonstrations
- [5] MAGIC criteria
- [2] Tableau Playtime?
- [5] homework prep
Basic stats

- mean
- median
- min
- max
- quartiles
- standard deviation
- normalization

Don't be scared by these terms. Most of statistics involves fancy ways of counting, and most folks are good at counting!
Correlation - how connected two variables are - doesn't imply that one causes the other. Use the linked-to 538 piece to have students in pairs play around with how easy it is to find correlations, or not find them, but picking and choosing your data inputs.
Have a students present a summary of the assigned readings.
Make sure they cover:

- What algorithmic techniques should we be aware of when analyzing large sets of data? and their limitations?
a taxonomy of types of stories

Let's think about the types of stories you can tell with data.
Story Archetypes

- Changes over Time
- Interesting Factoids
- Surprising Connections
- Personal Experiences
- Revealing Comparisons

This list is useful scaffolding for thinking about the kinds of stories you can find to tell. It isn’t a complete list by any means, but is useful to think through as you’re digging into data. I’ll give a concrete example of each.
In 2009 the city of Somerville, MA, USA did an audit of all of its trees. This data includes the exact location, species, health condition and more about each of the ~50,000 trees in the city. We'll use that data to think about the types of stories you can tell.
Humans experience the world through time, so before/after stories are naturally relatable.

This dot-map shows one area of town. Each dot is one tree, colored based on its health (red=poor, yellow-ok, green=good). Like most dot maps this suffers from the occlusion problem (where some dots sit on top of others, blocking them from view). However, we can see one street that stands out as having lots of healthy trees and not many unhealthy ones. Why?

With my knowledge of the neighborhood, I know that this is Somerville Ave., which just before this data collection effort was repaved and redone to beautify it. This story demonstrates a change-over-time narrative. In this case it could be used to discuss the role of city infrastructure planning in creating a healthier treescape.
Interesting Factoids

Sometimes the odd (aka outlier) datapoint is the one to dig into for a story.

We often try to start examining data by looking at it in aggregate. Don't forget that sometimes the solitary story of one data point can be the most compelling. This is a picture of a Hazelnut tree. There is only one in the whole city of Somerville. How did it get there?

This kind of story is an interesting "Factoid" - a single data point that can be the spark of a longer story.
People often look for connections between different parts of the data - in statistics there are called correlations. We talked about these before, where I cautioned that a correlation doesn't imply that one thing causes another.

Here is a chart of the overall satisfaction of residents in each part of town. One could wonder if there is a connection between the tree density in one part of town and the overall happiness.
Another type of story to tell involves personal experiences. Here I've highlighted one street in town. All the dots shown are spruce trees - this one street has all the spruce trees in town! You could connect to people on that street by talking about how unique their street is, and maybe even spark a campaign to rename their street to "Spruce St"!
The last type of story in my list is the revealing comparison. The idea here is to compare one slice of the data to another slice.

Here I'm showing two bubble charts. The left shows the population of each species of tree in town. You can see Maple and Pear are the most common. On the right I'm showing the population of trees just along Somerville Ave (one of the main streets). They look similar, and if you do the counting you'll find Somerville Ave to be the most similar to the overall population of trees in town. So if someone wanted to visit and see the tree leaves changing color in fall, you could point them towards Somerville Ave as a reasonable representation of the whole town.
Let's have some fun!
databasic.io/wordcounter

- team up
- find a story in some lyrics
- grab some paper and crayons
- sketch a visual of your story

Run the [WordCounter Sketch-a-Story activity (PDF)](#).
Analysis

- Exploring quantitative data with Tableau (demo)
  - More stats-oriented people could try JMP or Minitab (but they don’t give much for free)
- Exploring large qualitative data with overview
- Data Science Toolkit virtual machine for API-level access
- OpenRefine for finding correlations
- Sentiment analysis is popular, but has limitations
- Qualitative text analysis? (nineteen is promising)

There are lots of tools to help you here.
Abelson's MAGIC

- **Magnitude** - size of the claim
- **Articulation** - how precise is your claim
- **Generality** - is it valid in multiple contexts
- **Interestingness** - can this change beliefs in a way that matters?
- **Credibility** - do you believe it?

Review Abelson's MAGIC criteria from the reading. Remember these are things to keep in mind while making your arguments.
homework

- read stuff
- grad student to present reading on visual narratives