




Data Storytelling Studio

getting & cleaning data

CMS.631/831
Rahul Bhargava



Prep: setup room in circle, do nametags

Agenda

- [10] Review data logs
- [10] Getting data
- [10] Grad student presentation on open data papers
- [20] Cleaning data
- [10] Presentation crit
- [5] Homework prep

data log pair & share

the most nefarious?

the most benign?

the most surprising?

Students should have tracked all the types of data they create for one 24-hour period. Pair them up and have them spend 5 to 10 minutes talk about these three questions. Then bring them all back together and ask them to share anything interesting they talked about. Be sure to highlight any data that they didn't realize was being collected, any data they got worried about, and even mention potential uses/misuses.

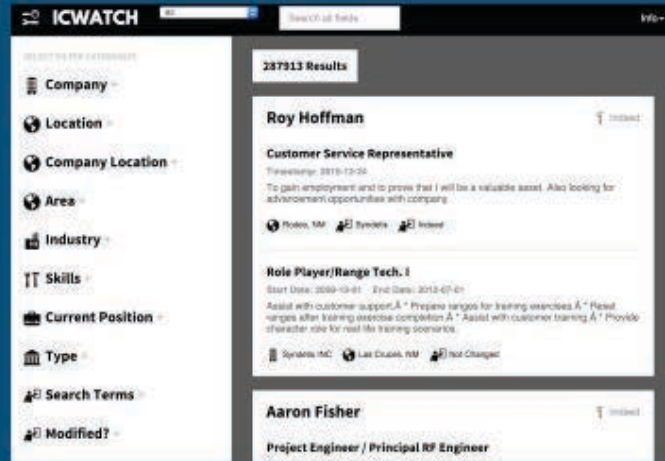
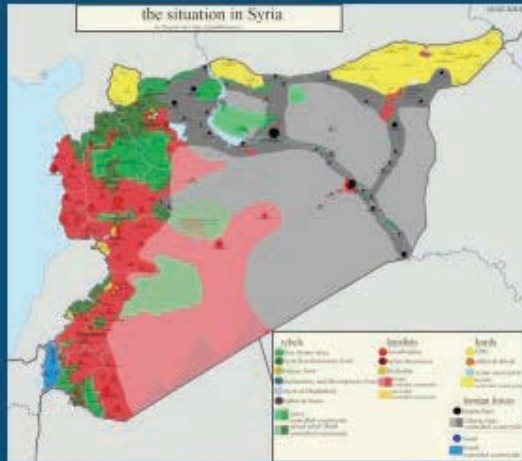
Getting data

Sources of Data

- Official sources (ic. govt agency)
- Advocacy / interest groups
- Personal knowledge
- Make it yourself

Rembering the "Asking Good Questions" WTFCSV activity, we have a list of sources for getting data

If the data doesn't exist?



Left © Syrian Civil War Map; right © ICWATCH. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>

If the data doesn't exist you can make it yourself.

Early on in the conflict in Syria a Dutch teenager scoured videos and message boards to create the authoritative source for maps of who controlled which territory. The example of the left shows one of these maps. Scouring public sources and aggregating information let an isolated individual to become the relied-on standard across the globe.

The example on the right comes from ICWatch - a database of people who list themselves on LinkedIn as working within the US Intelligence Community. Creating this database, from publicly posted data, revealed far more than was previously known about classified projects within the defense and intelligence sector.

Tools to Scrape



Here's a chart I like to draw to map out a space of tools. The vertical axis measure how easy to learn a tool is. The horizontal axis measure how many things a tool does. The top right quadrant is what I call the "iMovie" corner - for tools that do one thing and are pretty easy to learn. The bottom left is where computer programming goes; you want to avoid it at all costs because it is hard and takes a long time.

This particular chart shows some tools people use to scrape data.

open data papers

Joel Gurin. 2014. Open Governments, Open Data: A New Lever for Transparency, Citizen Engagement, and Economic Growth. SAIS Review of International Affairs 34, 1 (2014), 71–82.

Michael B. Gurstein. 2011. Open data: Empowering the empowered or effective data use for everyone? First Monday 16, 2 (January 2011).

Reading discussion questions:

- Access/interpretation/use - is Gurstein's model for "effective use" too technology centric and thus self-defeating?
- What models of citizenship are they invoking? Does the "Monitorial" definition of citizenship feel realistic or idealistic to you?
- Remember that open data is the beginning, not the end of the process of citizen engagement
- Social context is critical to think about data and use (entrenching and service existing power structures)
- Gurin argues for the economic impact of open data, and has numerous private corporate examples, but after Burghart do you trust these?

How have you seen data stored?

Once we have the data, we need to think about how we should store it to make it easy to use!

Storage strategies

- .csv files
- relational databases
- non-relational databases
- text files
- .pdf files
- HTML tables

- XLS / CSV files are the standard for most organizations
- Relational databases help when you have lots of data and it all relates to one another. For instance perhaps you have many sites within a food rescue organization, all of whom receive many different types of donations from multiple sources. The sites, donations, and sources would all be things that are related to each other.
- Non-relational databases are good when you need to store and access less formally structured information. They're very trendy right now.
- Text files are usually more unstructured, and can be harder to work with unless the text is all qualitative.
- PDF files are the worst, and sadly are a standard for open data releases from governments that don't really want you to have the data.
- HTML tables are nice, but sometimes hard to turn into something you can use in an analysis tool like Tableau or Excel.

What is "clean" data?

Ask folks what they think "clean data" means. Write up points on the board for a few minutes.

Clean Data

- **Consistency:** are observations always entered the same?
- **Completeness:** do you have coverage of the topic?
- **Usability:** machine readability?
- **Atomicity:** row-based normalization

Since we use machines to operate on data, machine-readability is a strong criteria.

And don't forget about the metadata!

See [the Quartz guide to bad data](#)

Here's what "clean data" means to me. Remember that 80% of the time of any data project is usually spent cleaning the data, so this is a big deal.

Remember that the metadata is important too - this describes what the data is. What does each column header mean? When was the data collected? What do blanks or zeros mean? What was left out (intentionally or unintentionally)

About "Tidy" Data

Hadley Wickham. 2014. Tidy Data. *Journal of Statistical Software* 59, 10 (August 2014).

Reading review:

- Run through one of Wickham's comparisons of row-atomic vs. not

Tools to clean



Here's a chart I like to draw to map out a space of tools. The vertical axis measure how easy to learn a tool is. The horizontal axis measure how many things a tool does. The top right quadrant is what I call the "iMovie" corner - for tools that do one thing and are pretty easy to learn. The bottom left is where computer programming goes; you want to avoid it at all costs because it is hard and takes a long time.

This particular chart shows some tools people use to clean data. Don't forget the power of find and replace!

Cleaning geographic data

Geoparsing: finding references to geographic places in text

tricky, but my [Cliff tool](#) does some of this

Geocoding: turning an address into latitude/longitude coordinates

[BatchGeo](#) can do a lot for you for free

Getting data out of PDF files

assuming the text is readable:

Let's open up [an example PDF](#) and try out [Tabula](#) (the best I've seen so far)

If you're a programmer, [pdftables](#) is a useful option

if it is an image:

you're in trouble - the automated OCR toolchain isn't great

Cleaning text / numbers

misspellings? try [OpenRefine](#) to [cluster them](#)

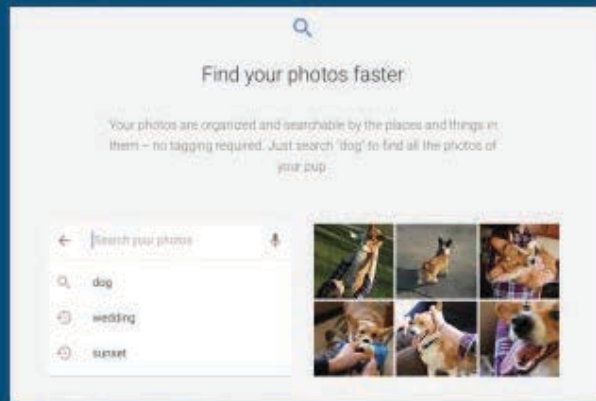
extracting data? try [regular expressions](#) ([use a cheatsheet](#)) ([learn it yourself](#))

splitting columns? remember [Excel can do some of this](#)

anonymizing? [scrubadub.io](#) is an in-progress tool to help

Using image data

You can analyze images qualitatively and quantitatively by repurposing tools like Google Photos.



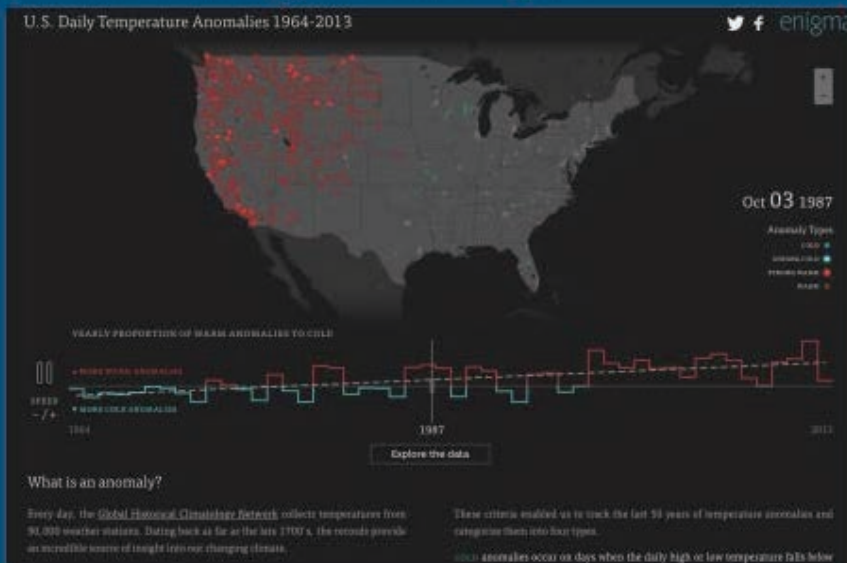
© Google. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>

Tools like Google Photos automatically detect objects in your photos now. You can repurpose that to drive your data analysis if you have lots of photos you are working with. It can answer questions like "how many photos in my set have a car" /

Another critique

Lets run another critique, this time of an interactive data visualization.

A more complex example



bit.ly/climate123

Ask folks these questions to help them read the chart and start to critique it.

- What datasets are being represented here?
 - US geography
 - Daily temperature "anomalies" (date, temperature, location)
 - Proportion of warm to cold anomalies
- What visual mappings are used to show data?
 - Map shows where the occurred
 - Colored dots shows temp of anomaly
 - Line chart shows ratio over time
- What is the one-sentence story here?
 - There are more and more warm temperature anomalies over time
- Do you think this story is well told?

homework

- install Tableau
- read stuff
- grad student to present reading on machine learning & big data

Tableau gives out free year-long licenses to students.

MIT OpenCourseWare
<https://ocw.mit.edu/>

CMS.631 Data Storytelling Studio: Climate Change
Spring 2017

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.