Old Dominion University

CS 795/895
Applied Visual Analytics
Spring 2013

Data (Formats, Gathering, Cleaning)

Dr. Michele C. Weigle

http://www.cs.odu.edu/~mweigle/CS795-S13/

Outline

- Data Formats
- Gathering Data
- Cleaning Data
- Resources
  - John Stasko's CS 7450 @ GaTech, Spring 2011
  - *Visualize This*, Ch 2
  - "Scraping for Journalism: A Guide for Collecting Data" - http://www.propublica.org/nerds/item/doc-dollars-guides-collecting-the-data
  - *Beautiful Data*, Ch 20

Data Formats

- Delimited Text
  - tabbed delimited
  - comma delimited (CSV)
- Extensible Markup Language (XML)
  - looks kinda like HTML
  - user-defined tags to identify data
- Javascript Object Notation (JSON)
  - collection of name/value pairs
  - smaller than XML
  - easier to parse

Data Formats
Example - Weather Data

CSV

```
20090101,26
20090102,34
20090103,27
```

XML

```
<weather_data>
  <observation>
        <date>20090101</date>
        <max_temp>26</max_temp>
  </observation>
  <observation>
        <date>20090102</date>
        <max_temp>34</max_temp>
  </observation>
  <observation>
        <date>20090103</date>
        <max_temp>27</max_temp>
  </observation>
</weather_data>
```

JSON

```
{"observations": [
   {"date":"20090101", "max_temp":26},
   {"date":"20090102", "max_temp":34},
   {"date":"20090103", "max_temp":27}
 ]
}
```

Data Formats
Converting Using Code

- Write a program to convert from one format to another

- awk - my favorite, but I'm old school
- Python
- Ruby
- Perl
- PHP

Data Formats
Converting Using Tools

- Just search Google for "csv to json", "csv to xml", "xml to json"

- Mr. Data Converter
  - http://shancarter.com/data_converter/
  - developed by a graphics editor at *The New York Times*
  - input: CSV or tab-delimited data
  - output: HTML table, JSON, MySQL, Python, PHP, Ruby, XML, ...

Gathering Data

- Data to be visualized can come from a variety of sources and in a variety of formats
  - our own experiments/research
  - provided by others (libraries, data warehouses)
  - web content

Gathering Data
Web Scraping

- What if your data is online, but isn't given in a nice Excel file?
- You'll need to scrape it from the web site

Gathering Data
Web Scraping

- Find out what input a website expects
  - fill in form and inspect the URI generated
  - experiment with changing URI parameters to understand the request format
- Understand the format of its response
  - generally sites will pull data from a database and present use the same template for display
  - useful tools
    - view source
    - Firefox's Firebug extension
    - Google Chrome's Inspect Element option (right-click)

http://www.propublica.org/nerds/item/scraping-websites

Web Scraping
Basic Process

- Identify the patterns
  - URLs, web page layout
- Iterate
  - visit each web page and grab the relevant data
  - if there are a lot of pages, this might take some time
- Store the data

Web Scraping
Example

- Let's grab data from the Weather Underground
  - http://wunderground.com
- Want temperature data for Norfolk for 2011
  - starting point: http://www.wunderground.com/history/airport/ORF/2011/1/1/DailyHistory.html
- Tools
  - Python
  - Beautiful Soup - Python script for reading web pages (http://www.crummy.com/software/BeautifulSoup)

http://weiglevm.cs.odu.edu/~mweigle/cs795s13/scraping-example.html

Web Scraping
Some Comments

- Some sites limit the rate of web requests you can make from a single IP
  - e.g. - Google Books API: 1,000 requests/day
- Sometimes the web site structure is difficult to parse
  - might want to download the HTML pages to a local disk before parsing
    - avoids re-downloading pages if the parsing breaks or while debugging the parsing script
- When storing the data you've collected, think about how you might use it later
  - put it in a usable format

Gathering Data
From Non-HTML Web Pages

- Flash
  - http://www.propublica.org/nerds/item/reading-flash-data
  - use browser tools to analyze what files are being loaded
  - clear cache, load Flash file (swf), look for data files being transferred (xml, json)
- PDF
  - http://www.propublica.org/nerds/item/turning-pdfs-to-text-doc-dollars-guide
  - use tools such as Adobe Acrobat Pro (PDF to HTML), pdftotext (PDF to plain text), or third-party sites (like http://www.cometdocs.com)
  - requires lots of cleaning

Cleaning Data

- Data in the real world is never in the form that you need
- Before you can visualize something, you have to get the data into a form that you can work with
- Data can be missing, have typos, be inconsistent, spread over multiple tables
- Two big issues:
  - format
  - accuracy

Data Cleaning
Tools

Quick Tools

- Data Science Toolkit
  - http://www.datasciencetoolkit.org/
  - lots of quick conversion tools
- Mr. People
  - http://people.ericson.net/
  - formats lists of names
- Mr. Data Converter
  - http://shancarter.com/data_converter/

Full Apps

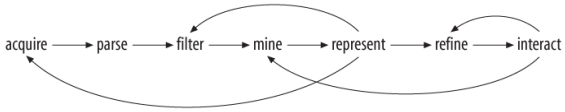
- Data Wrangler
  - http://vis.stanford.edu/wrangler/app/
  - video: http://vimeo.com/19185801
- Google Refine / Open Refine
  - http://code.google.com/p/google-refine/
  - video: http://www.youtube.com/watch?v=yNccGtn3Wb0&feature=player_embedded
  - more info: http://www.propublica.org/nerds/item/using-google-refine-for-data-cleaning

Data Cleaning
Accuracy is Essential

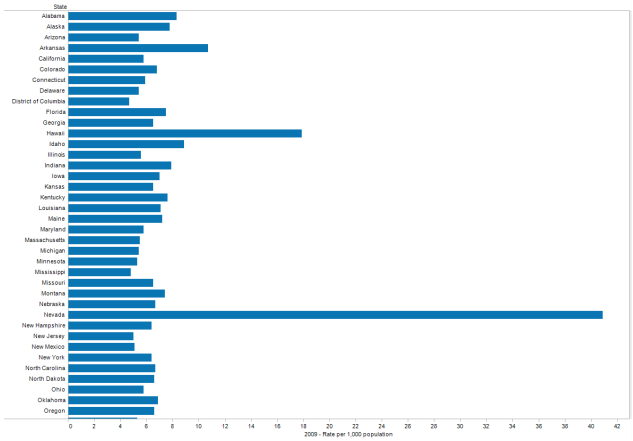
- *Must have accurate data before can trust the visualization*
- Nathan Yau was intern at *The New York Times*
  - one day, his entire goal was to verify 3 numbers in a dataset

Recall

Info Vis Stages Example

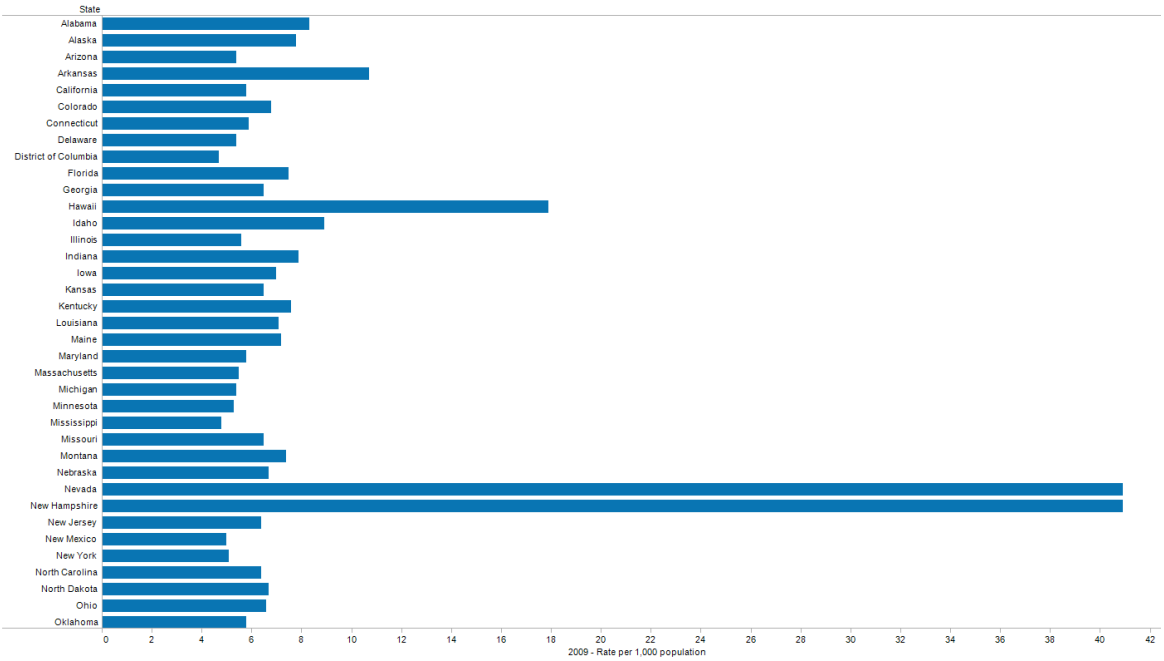


- Represent
  - Load filtered data into Tableau Public
  - View as map – *not exactly what I want*
  - View as bar chart



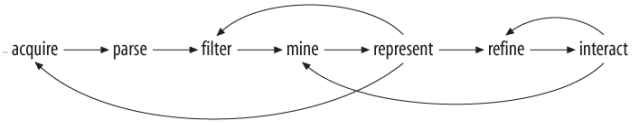
What the Chart First Looked Like

What?!?!



Seven Stages of Data Visualization

Back to Acquire…

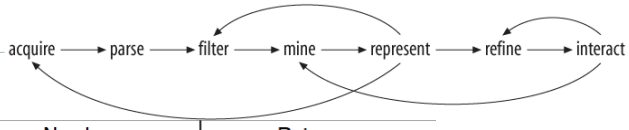


Let's examine the Excel file

	Rate 1,000 population \2						
	1990	2000	2005	2006	2007	2008	2009
Maryland	9.7	5.8	6.9	6.6	6.5	6.0	5.8
Massachusetts	7.9	6.7	6.2	5.9	5.9	5.7	5.5
Michigan	8.2	6.8	6.0	5.9	5.9	5.5	5.4
Minnesota	7.7	6.9	5.9	6.0	6.0	5.5	5.3
Mississippi	9.4	7.8	5.8	5.7	5.7	5.2	4.8
Missouri	9.6	7.3	7.0	6.9	6.9	6.8	6.5
Montana	8.6	7.6	7.4	7.5	7.5	7.7	7.4
Nebraska	8.0	72.2	7.0	6.8	6.8	6.9	6.7
Nevada	99.0	9.4	57.8	52.6	52.6	43.1	40.9
New Hampshire	9.5	6.0	7.2	7.1	7.1	6.8	40.9
New Jersey	7.6	6.0	5.7	5.5	5.5	5.4	6.4
New Mexico \5	8.8	8.0	6.6	6.9	6.9	4.0	5.0
New York \5	8.6	7.1	6.8	6.8	6.8	6.5	5.1
North Carolina	7.8	8.2	7.3	7.3	7.3	7.0	6.4
North Dakota	7.5	7.2	6.9	6.8	6.8	6.7	6.7
Ohio	9.0	7.8	6.5	6.3	6.3	6.0	6.6
Oklahoma	10.6	(NA)	7.3	7.3	7.3	7.1	5.8
Oregon	8.9	7.6	7.3	7.2	7.2	6.9	6.9
Pennsylvania	7.1	6.0	5.8	5.7	5.7	5.6	5.3
Rhode Island	8.1	7.6	7.0	6.5	6.5	6.2	5.9

Seven Stages of Data Visualization

Back to Acquire…

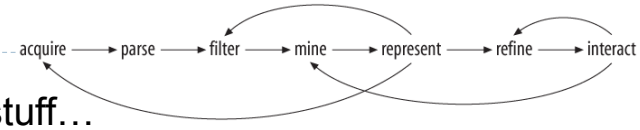


As compared to the PDF

	Number (1,000)			Rate per 1,000 population [2]		
	1990	2000	2009	1990	2000	2009
Maryland	46.3	40.0	32.4	9.7	7.5	5.8
Massachusetts	47.7	37.0	36.7	7.9	5.8	5.5
Michigan	76.1	66.4	53.1	8.2	6.7	5.4
Minnesota	33.7	33.4	28.4	7.7	6.8	5.3
Mississippi	24.3	19.7	14.5	9.4	6.9	4.8
Missouri	49.1	43.7	39.8	9.6	7.8	6.5
Montana	6.9	6.6	7.1	8.6	7.3	7.4
Nebraska	12.6	13.0	12.5	8.0	7.6	6.7
Nevada	120.6	144.3	108.2	99.0	72.2	40.9
New Hampshire	10.5	11.6	8.5	9.5	9.4	6.4
New Jersey	58.7	50.4	46.3	7.6	6.0	5.0
New Mexico [5]	13.3	14.5	10.2	8.8	8.0	5.1
New York [5]	154.8	162.0	120.1	8.6	7.1	6.4
North Carolina	51.9	65.6	65.8	7.8	8.2	6.7
North Dakota	4.8	4.6	4.3	7.5	7.2	6.6
Ohio	98.1	88.5	64.8	9.0	7.8	5.8
Oklahoma	33.2	15.6	23.5	10.6	(NA)	6.9
Oregon	25.3	26.0	23.5	8.9	7.6	6.6
Pennsylvania	84.9	73.2	64.2	7.1	6.0	5.3
Rhode Island	8.1	8.0	6.5	8.1	7.6	5.9

Seven Stages of Data Visualization

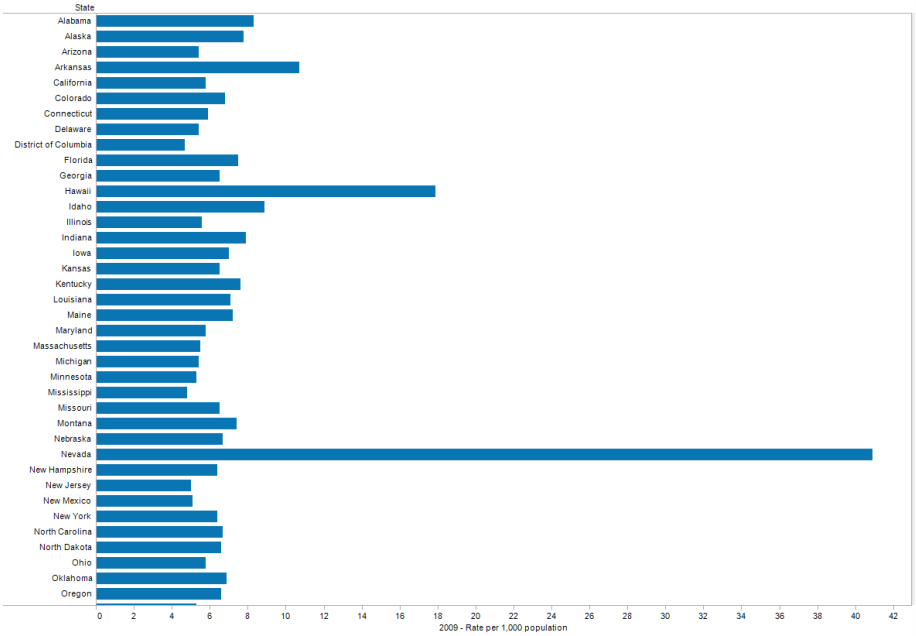
Back to Acquire…



But wait, there's more funny stuff…

	PDF	Excel
Maryland	5.8	5.8
Massachusetts	5.5	5.5
Michigan	5.4	5.4
Minnesota	5.3	5.3
Mississippi	4.8	4.8
Missouri	6.5	6.5
Montana	7.4	7.4
Nebraska	6.7	6.7
Nevada	40.9	40.9
New Hampshire	6.4	40.9
New Jersey	5.0	6.4
New Mexico [5]	5.1	5.0
New York [5]	6.4	5.1
North Carolina	6.7	6.4
North Dakota	6.6	6.7
Ohio	5.8	6.6
Oklahoma	6.9	5.8
Oregon	6.6	6.9
Pennsylvania	5.3	5.3
Rhode Island	5.9	5.9

Final Chart



So Far...

- Intro to Visual Analytics
- Intro to Info Vis
- Tools for Data Gathering and Data Cleaning

Coming Up:

- Graph Types and Design Principles
- PHP/MySQL