



Data Storytelling Studio

getting & cleaning data

CMS.631/831
Rahul Bhargava



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?

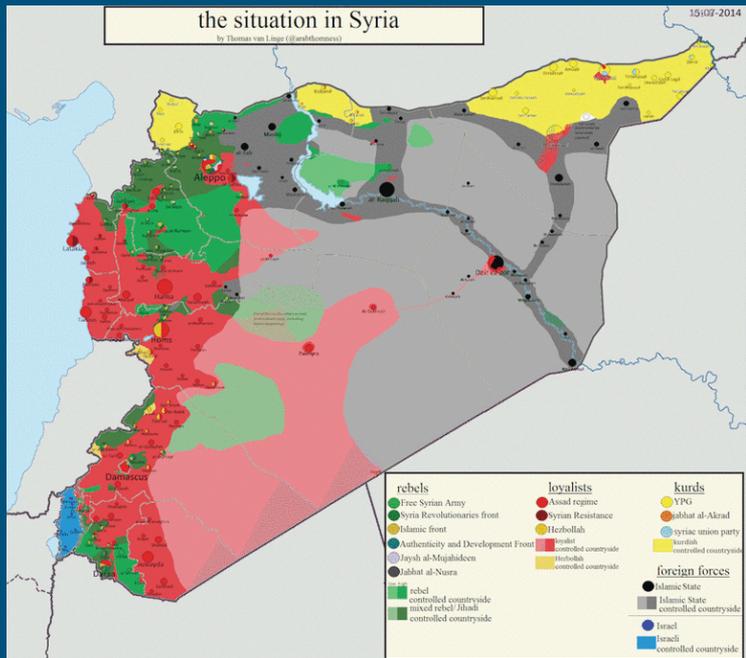


Getting data

Sources of Data

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

If the data doesn't exist?



<https://syriancivilwarmap.com/>

ICWATCH All Search all fields Info

SELECT FILTER CATEGORIES

- Company +
- Location +
- Company Location +
- Area +
- Industry +
- Skills +
- Current Position +
- Type +
- Search Terms +
- Modified? +

287913 Results

Roy Hoffman Indeed

Customer Service Representative
Timestamp: 2015-12-24
To gain employment and to prove that I will be a valuable asset. Also looking for advancement opportunities with company.

Rodeo, NM Syndetix Indeed

Role Player/Range Tech. I
Start Date: 2008-12-01 End Date: 2012-07-01
Assist with customer support. * Prepare ranges for training exercises. * Reset ranges after training exercise completion. * Assist with customer training. * Provide character role for real life training scenarios.

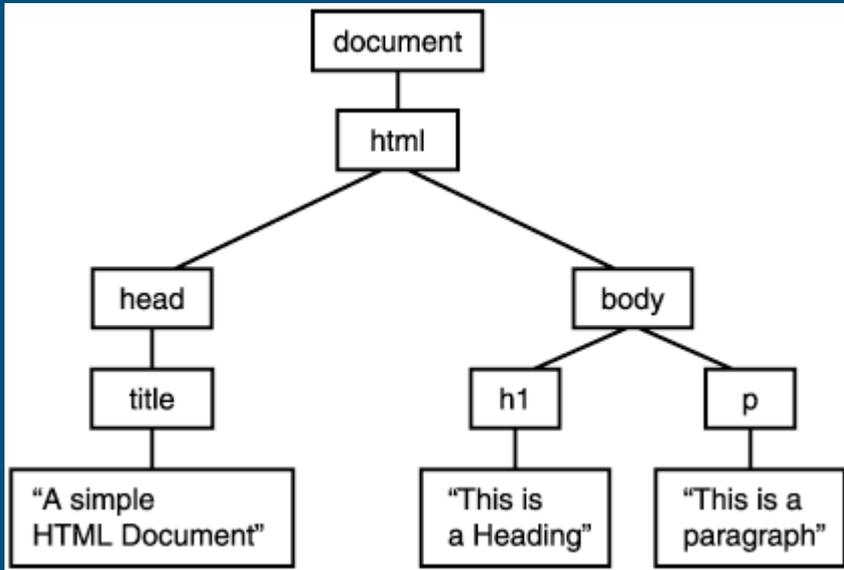
Synetix INC Las Cruces, NM Not Changed

Aaron Fisher Indeed

Project Engineer / Principal RF Engineer

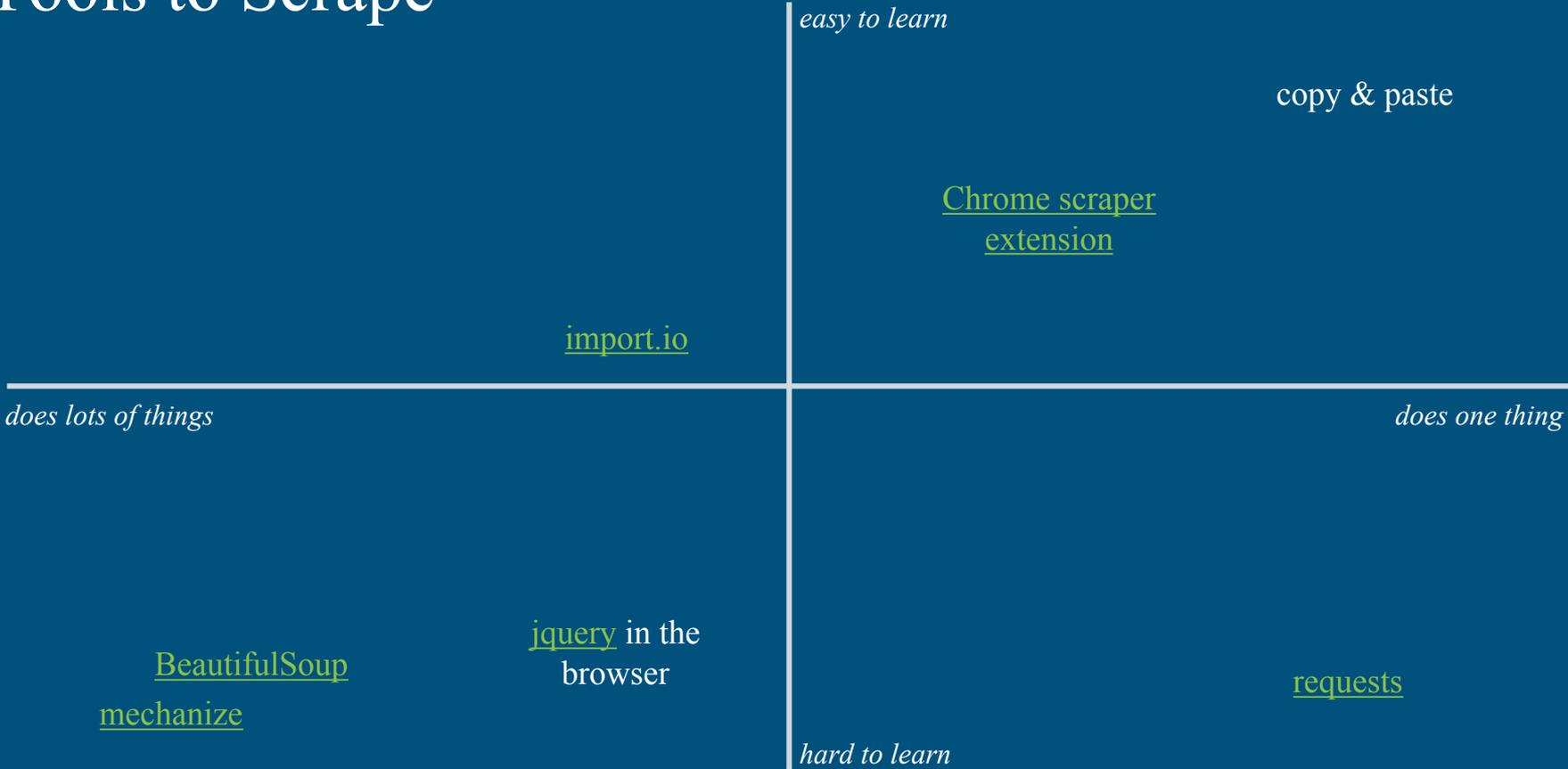
<https://icwatch.wikileaks.org/>

Document Object Model (ie. "the DOM")



```
<!DOCTYPE html>
<html class=" js no-touch localStorage applicationcache svg inlinesvg display-
table enhanced enhanced-rendering fontface" style>
  <head>...</head>
  <body itemscope itemtype="http://schema.org/WebPage" class="type-home logged-
out">
    <!-- empty statusBar -->
    <div id="contain">
      <!-- Page header -->
      <!-- topSection1 / -->
      <header class="bg-header clearfix">
        ::before
        <div class="bg-header_subinfo bg-header_subinfo--tablet">...</div>
        <div id="smartbar" class="bg-header__smartbar bg-header__smartbar--show"
data-hide-on-load="true">...</div>
        <div class="bg-header__menus">...</div>
        <div class="bg-header__subinfo">...</div>
        <div href class="bg-header__centerpiece bg-header__logo-wrapper">...</div>
        <nav class="bg-header__nav">
          <ul>...</ul>
        </nav>
        <nav class="bg-header__takeover" aria-hidden="true">...</nav>
        <aside class="bg-header__pop-up" aria-hidden="true">...</aside>
        ::after
      </header>
      <div class="bg-header__smartbar--veil bg-header__smartbar--veil-show">
...</div>
      <!-- Page header ends -->
      <div class="ad billboard" id="ad_outofpage2" data-adname-complete=
"true"></div>
    </body>
```

Tools to Scrape



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).

How have you seen
data stored?

Storage strategies

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

What is "clean" data?

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.

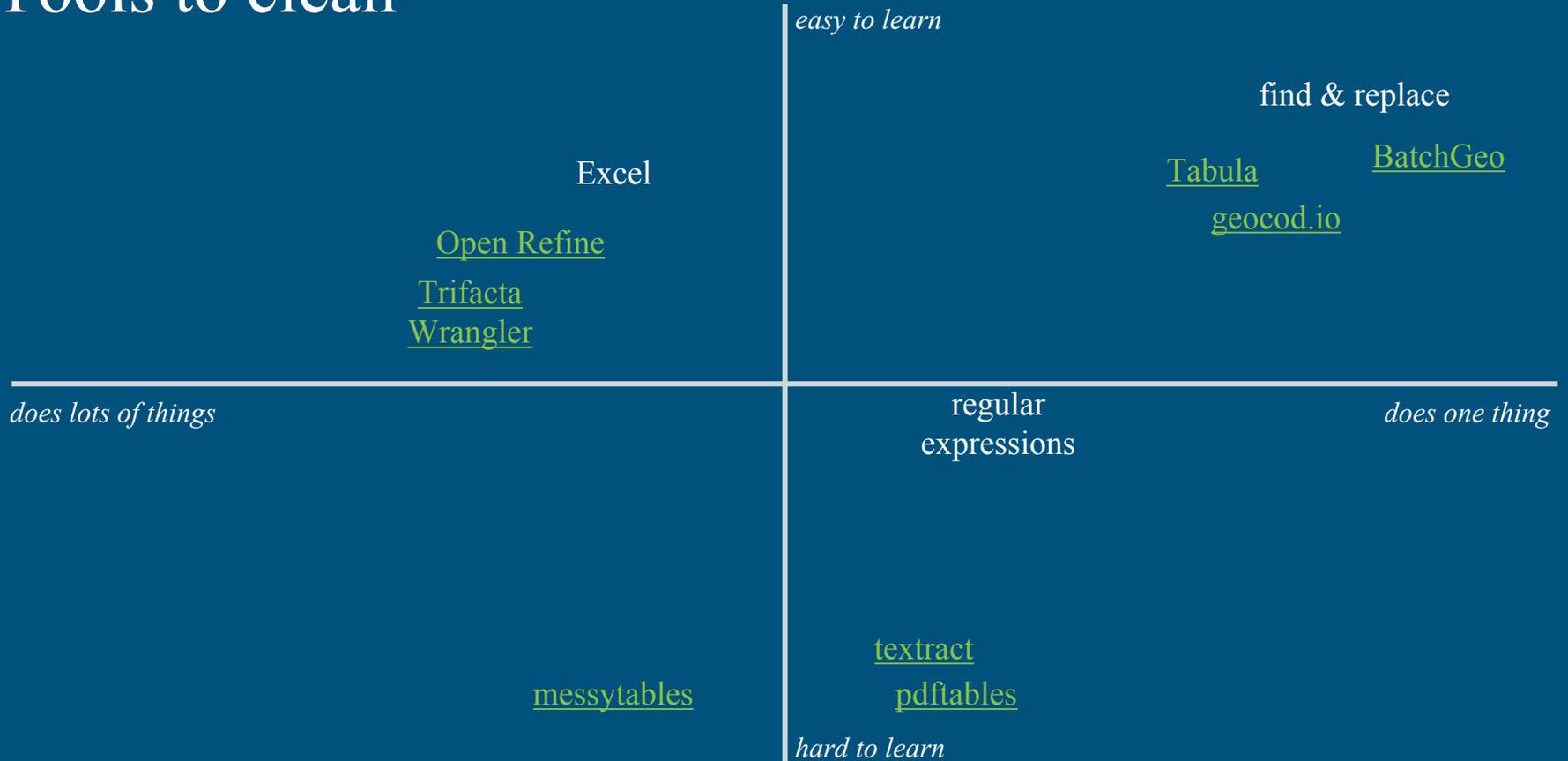
And don't forget about the metadata!

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

About "Tidy" Data

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

Tools to clean



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](#)

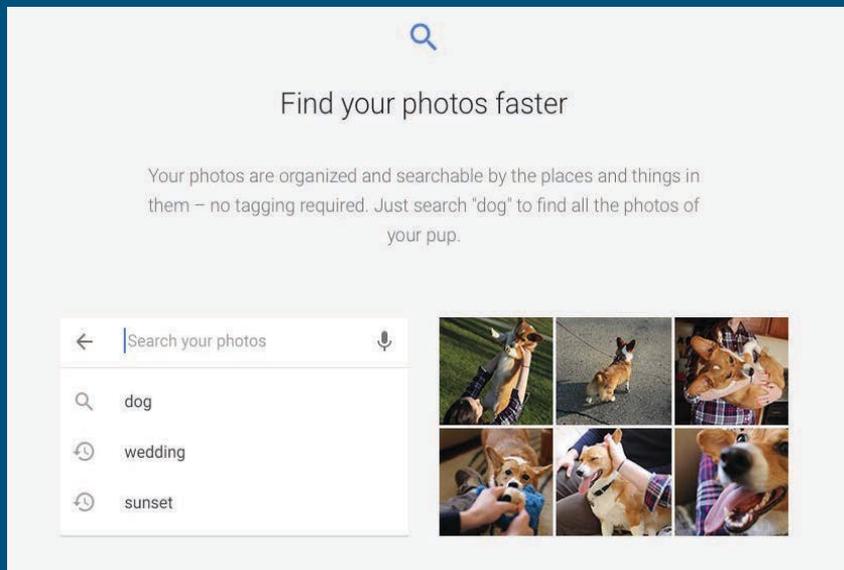
extracting data? try [regular expressions](#) ([use a cheatsheet](#)) ([learn it yourself](#))

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

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

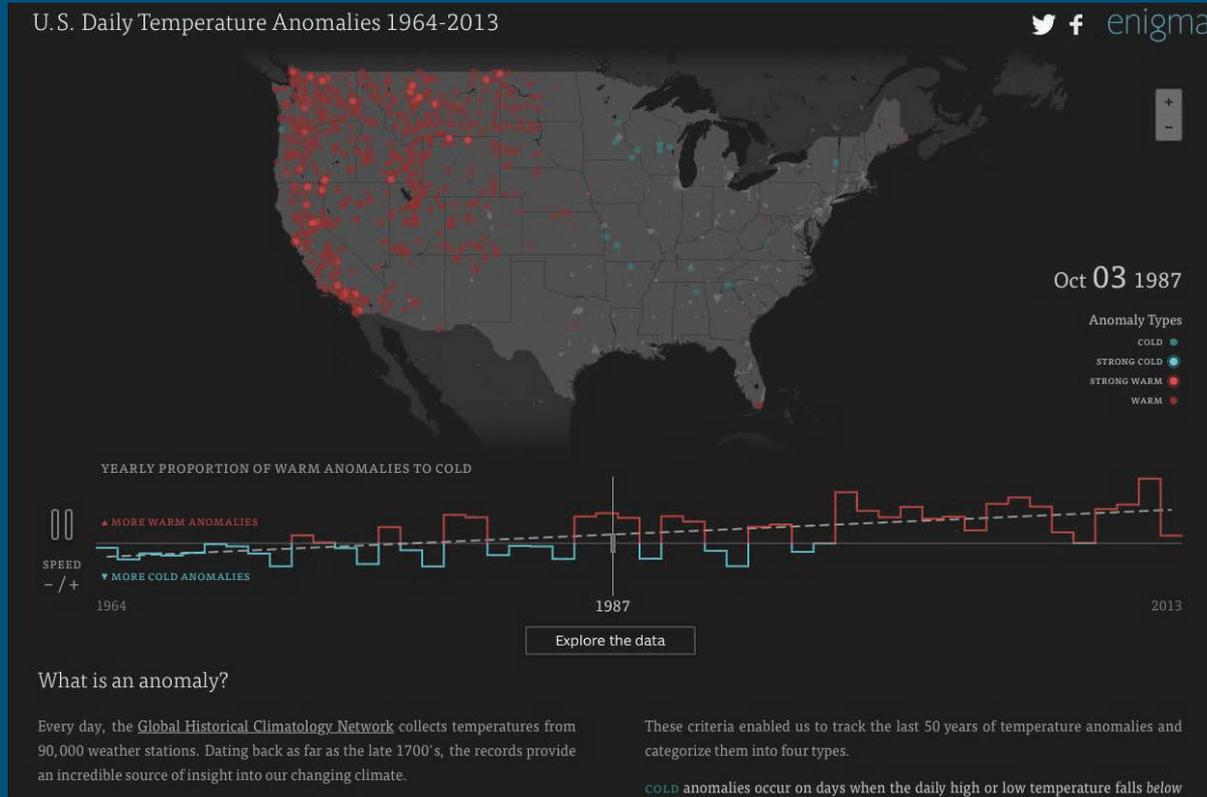
Using image data

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



Another critique

A more complex example



bit.ly/climate123

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

homework

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

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>.