Lecture 4: Data Collection and Munging

Instructor: Saravanan Thirumuruganathan

Outline

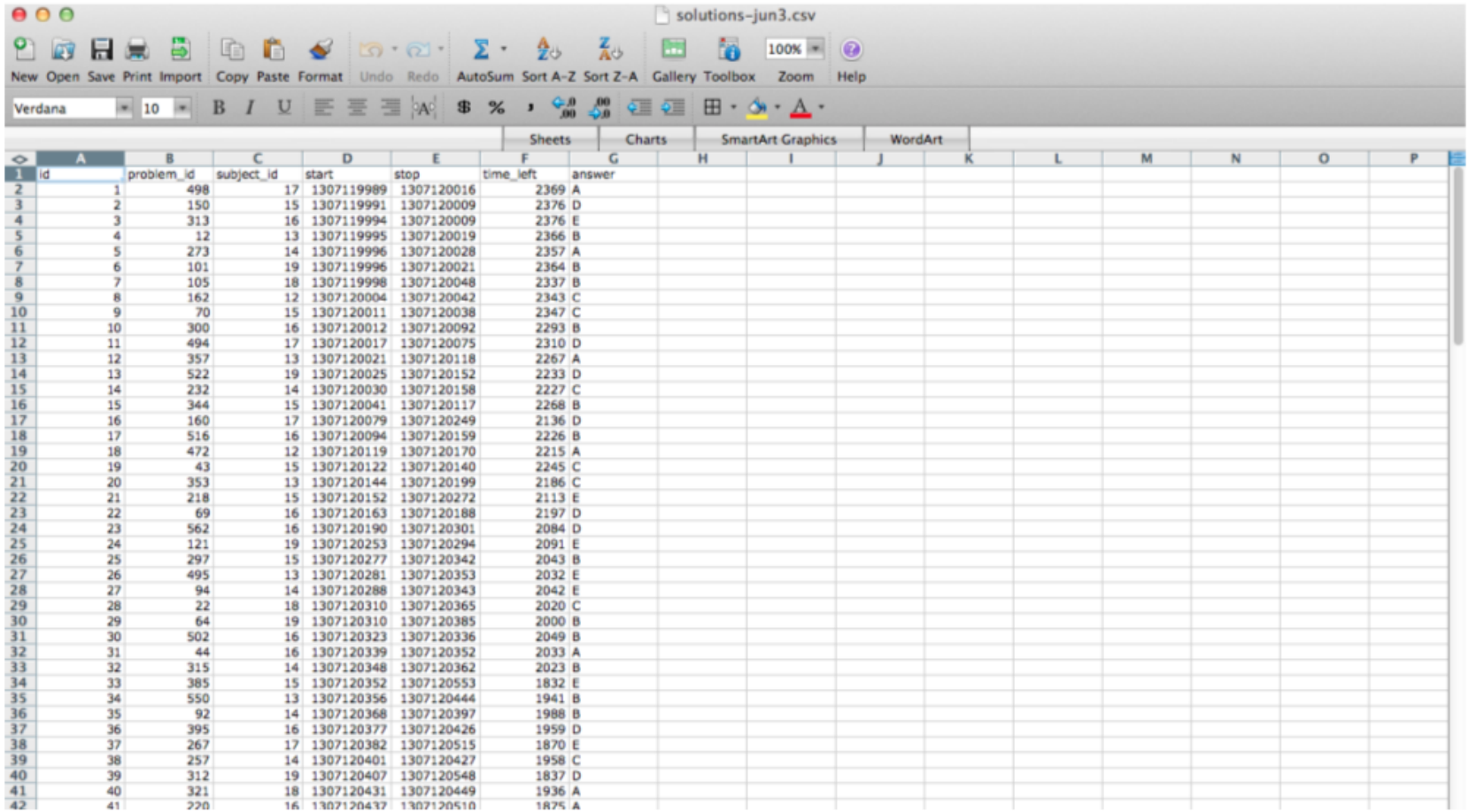
1. Data Collection and Scraping
2. Web Scraping basics

In-Class Quizzes

- **URL:** `http://m.socrative.com/`
- **Room Name:** **4f2bb99e**

Data Collection

What you wish data looked like?

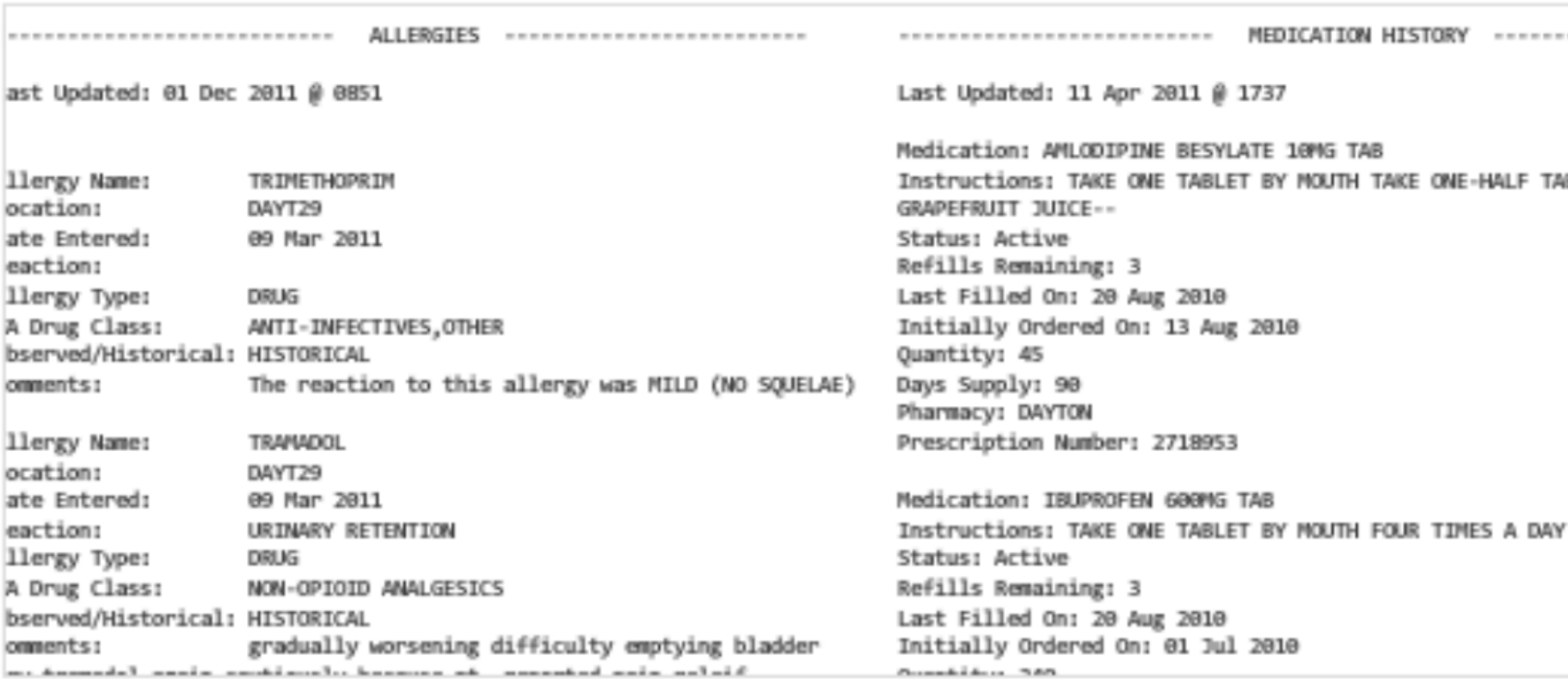


	A	B	C	D	E	F	G
1	id	problem_id	subject_id	start	stop	time_left	answer
2	1	498	17	1307119989	1307120016	2369	A
3	2	150	15	1307119991	1307120009	2376	D
4	3	313	16	1307119994	1307120009	2376	E
5	4	12	13	1307119995	1307120019	2366	B
6	5	273	14	1307119996	1307120028	2357	A
7	6	101	19	1307119996	1307120021	2364	B
8	7	105	18	1307119998	1307120048	2337	B
9	8	162	12	1307120004	1307120042	2343	C
10	9	70	15	1307120011	1307120038	2347	C
11	10	300	16	1307120012	1307120092	2293	B
12	11	494	17	1307120017	1307120075	2310	D
13	12	357	13	1307120021	1307120118	2267	A
14	13	522	19	1307120025	1307120152	2233	D
15	14	232	14	1307120030	1307120158	2227	C
16	15	344	15	1307120041	1307120117	2268	B
17	16	160	17	1307120079	1307120249	2136	D
18	17	516	16	1307120094	1307120159	2226	B
19	18	472	12	1307120119	1307120170	2215	A
20	19	43	15	1307120122	1307120140	2245	C
21	20	353	13	1307120144	1307120199	2186	C
22	21	218	15	1307120152	1307120272	2113	E
23	22	69	16	1307120163	1307120188	2197	D
24	23	562	16	1307120190	1307120301	2084	D
25	24	121	19	1307120253	1307120294	2091	E
26	25	297	15	1307120277	1307120342	2043	B
27	26	495	13	1307120281	1307120353	2032	E
28	27	94	14	1307120288	1307120343	2042	E
29	28	22	18	1307120310	1307120365	2020	C
30	29	64	19	1307120310	1307120385	2000	B
31	30	502	16	1307120323	1307120336	2049	B
32	31	44	16	1307120339	1307120352	2033	A
33	32	315	14	1307120348	1307120362	2023	B
34	33	385	15	1307120352	1307120553	1832	E
35	34	550	13	1307120356	1307120444	1941	B
36	35	92	14	1307120368	1307120397	1988	B
37	36	395	16	1307120377	1307120426	1959	D
38	37	267	17	1307120382	1307120515	1870	E
39	38	257	14	1307120401	1307120427	1958	C
40	39	312	19	1307120407	1307120548	1837	D
41	40	321	18	1307120431	1307120449	1936	A
42	41	220	16	1307120437	1307120510	1875	A

What does data really look like?

```
@HWI-EAS121:4:100:1783:550#0/1
CGTTACGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACGGATCTCGTATGCGGTCTGCTGCGTGACAAGACAGGGG
+HWI-EAS121:4:100:1783:550#0/1
aaaaa`b_aa`aa`YaX]aZ`aZM^Z]YRa]YSG[[ZREQLHESDHNDDHNMEEDDMPENITKFLFEEDDDHEJQMEDDD
@HWI-EAS121:4:100:1783:1611#0/1
GGGTGGGCATTTCCACTCGCAGTATGGGTTGCCGCACGACAGGCAGCGGTCAGCCTGCGCTTTGGCCTGGCCTTCGGAAA
+HWI-EAS121:4:100:1783:1611#0/1
a```^\__`_```^a``a`^a_^__]a_]\]`a_______`_^^`]X]_]XTV_\]]NX_XVX]]_TTTTG[VTHPN]VFDZ
@HWI-EAS121:4:100:1783:322#0/1
CGTTTATGTTTTTGAATATGTCTTATCTTAACGGTTATATTTTAGATGTTGGTCTTATTCTAACGGTCATATATTTTCTA
+HWI-EAS121:4:100:1783:322#0/1
abaa`^aaaaabbbaabaabbbbbb`bbbb_bbbbbbbb`bbbaV^_a``a``]``aT]a__V\]]_]^a`]a_abbaV__
@HWI-EAS121:4:100:1783:1394#0/1
GGGTCTTTATTGGTCTGGTGATCCCCCATATTCTCCGGTTGTGTGGTTTAACCGATCATCGCGCATTACTTCCCGGCTGC
+HWI-EAS121:4:100:1783:1394#0/1
```[aa\b^^[]aabbb][`a_abbb`a``bbbbbabaabaaaab_VZa_^___bab_X`[a\HV_[_]_[^_X\T_VQQ
@HWI-EAS121:4:100:1783:207#0/1
CCCTGGGAGATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCTCGTATGCCGTCTTCTGCTTGAAAAAAAAAACA
+HWI-EAS121:4:100:1783:207#0/1
abba`Xa\^\\`aa]ba__bba[a_O_a`aa`aa`a]^V]X_a^YS\R_\H_[]\ZTDUZZUSOPX]]POP\GS\WSHHD
@HWI-EAS121:4:100:1783:455#0/1
GGGTAATTCAGGGACAATGTAATGGCTGCACAAAAAAATACATCTTTCATGTTCCATTGCACCATTGACAAATACATATT
+HWI-EAS121:4:100:1783:455#0/1
abb_babbabaabbbbbbbbbbbbbbba\`b`\abbbabbbbabbbbbbaabbbbb`bb`ab_O_bab_Q_bbabaa_a
```
```

What does data really look like?



---------- ALLERGIES ----------

ast Updated: 01 Dec 2011 @ 0851

llergy Name: TRIMETHOPRIM
ocation: DAYT29
ate Entered: 09 Mar 2011
eaction:
llergy Type: DRUG
A Drug Class: ANTI-INFECTIVES,OTHER
bserved/Historical: HISTORICAL
omments: The reaction to this allergy was MILD (NO SQUELAE)

llergy Name: TRAMADOL
ocation: DAYT29
ate Entered: 09 Mar 2011
eaction: URINARY RETENTION
llergy Type: DRUG
A Drug Class: NON-OPIOID ANALGESICS
bserved/Historical: HISTORICAL
omments: gradually worsening difficulty emptying bladder

---------- MEDICATION HISTORY ----------

Last Updated: 11 Apr 2011 @ 1737

Medication: AMLODIPINE BESYLATE 10MG TAB
Instructions: TAKE ONE TABLET BY MOUTH TAKE ONE-HALF TA
GRAPEFRUIT JUICE--
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 13 Aug 2010
Quantity: 45
Days Supply: 90
Pharmacy: DAYTON
Prescription Number: 2718953

Medication: IBUPROFEN 600MG TAB
Instructions: TAKE ONE TABLET BY MOUTH FOUR TIMES A DAY
Status: Active
Refills Remaining: 3
Last Filled On: 20 Aug 2010
Initially Ordered On: 01 Jul 2010

Big Data Borat
@BigDataBorat

Follow

In Data Science, 80% of time spent prepare data, 20% of time spent complain about need for prepare data.

RETWEETS 380 FAVORITES 129

6:47 PM - 26 Feb 2013

Oliver Mason @ojmason · 27 Feb 2013
@BigDataBorat Are you sure it's not the other way round?

What Do Analysts Do?

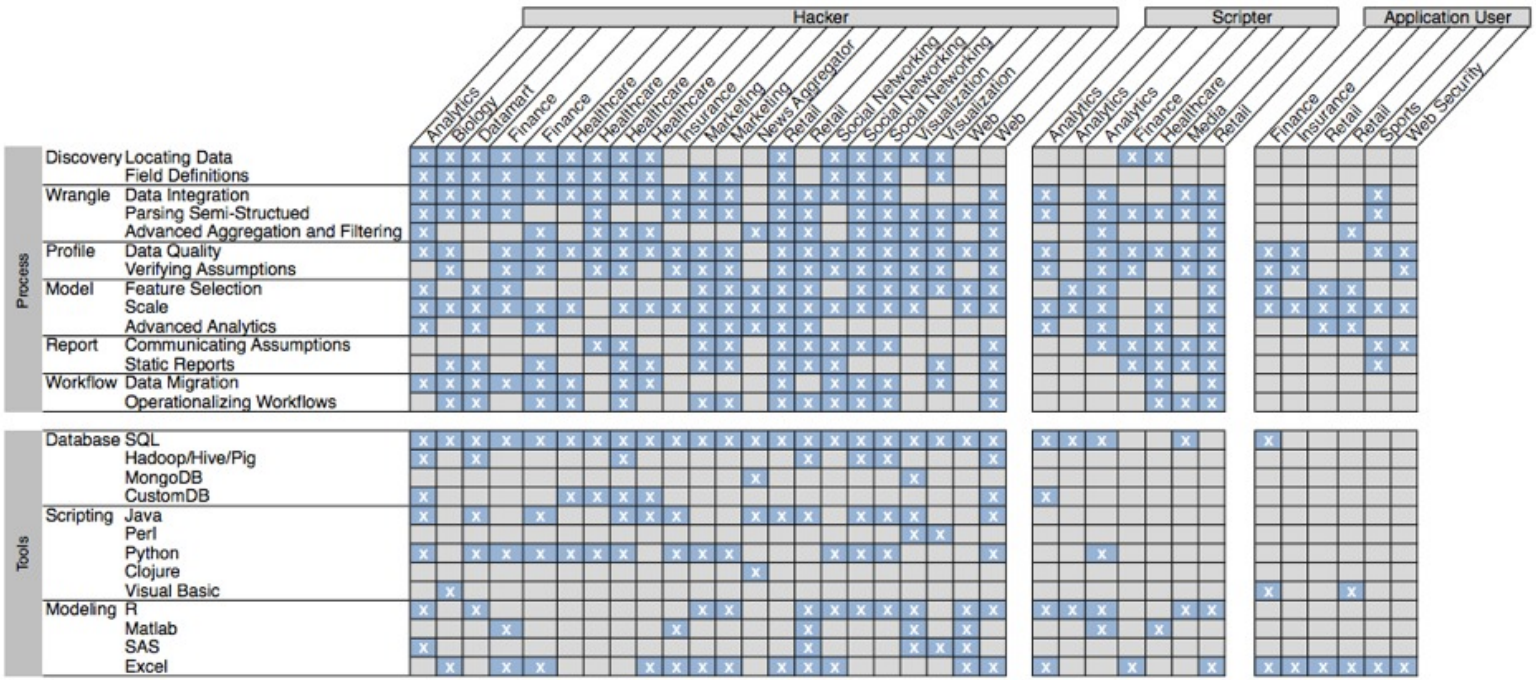


Fig. 1. Respondents, Challenges and Tools. The matrix displays interviewees (grouped by archetype and sector) and their corresponding challenges and tools. *Hackers* faced the most diverse set of challenges, corresponding to the diversity of their workflows and toolset. *Application users* and *scripters* typically relied on the IT team to perform certain tasks and therefore did not perceive them as challenges.

What Do Analysts Do?

I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any analysis. Most of the time once you transform the data you just do an average... the insights can be scarily obvious. It's fun when you get to do something somewhat analytical.

Data Collection Ethics

- OK: Public, non-sensitive, anonymized, fully referenced information (always cite sources)
- If in doubt, don't!
- Be a good web citizen:
  - Honor robots.txt
  - Obey rate limits. Do not overload servers
  - Obey relevant copyright/license restrictions
  - Know about fair use and its restrictions

Data Collection: 4 Things You Should Have

1. Raw data
2. Tidy data set
3. A code/process book describing each variable and its values in tidy data set
4. A explicit and exact recipe you used to go from 1 to 2, 3.

Tidy Data

1. Each variable you measure should be in one column
2. Each different observation of that variable should be in a different row
3. There should be one table for each "kind" of variable
4. If you have multiple tables, they should include a column in the table that allows them to be linked

Other Tips:

- Include a row at top of each file with variable names
- Make variable names human readable AgeAtDiagnosis instead of AgeDx
- In general, data should be saved in one file per table

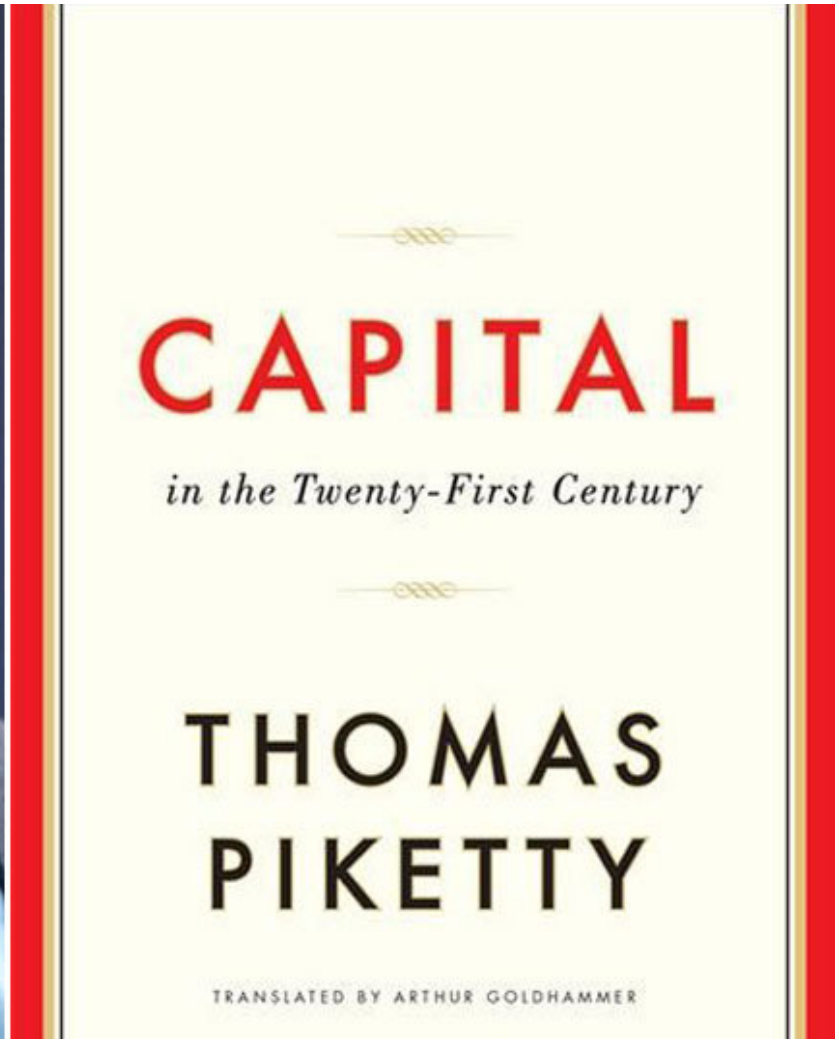
Why is the instruction list important?

Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff

Thomas Herndon* Michael Ash Robert Pollin

April 15, 2013

Why is the instruction list important?[1]



[1] http://telegraph.co.uk

Data Access Schemes

- **Bulk downloads:** Wikipedia, IMDB, Million Song Database, etc. See list of data web sites on the Resources page
- **API access:** NY Times, Twitter, Facebook, Foursquare, Google, ...
- **Web scraping:** For everything else

Data Formats

- Delimited Values
  - Comma Separated Values (CSV)
  - Tab Separated Values (TSV)
- Markup Languages
  - Hypertext Markup Language (HTML / XML)
  - JavaScript Object Notation (JSON)
  - Hierarchical Data Format (HDF5)
- Ad Hoc Formations
  - Graph edge lists, voting records, fixed width files, ...

JSON

- Light weight text-based interchange format
- Language independent
- Based on Javascript
- Very easy to use and manipulate
- All browsers and major languages have great support

JSON

- A collection of name/value pairs:
  $$\{\text{“}a\text{”} : 1, \text{“}b\text{”} : 2, \text{“}c\text{”} : 3, \text{“}d\text{”} : 4, \}$$
- An ordered list of values:
  $$[1, 2, 3, \text{“}blah\text{”}]$$
- Or a combination:
  $$\{\text{“}a\text{”} : [1, 2, 3, 4], \text{“}b\text{”} : [5, 6, 7, 8], \text{“}c\text{”} : 4\}$$

XML

- Tree structured

  Multiple field values, complex structure

- "Self-describing": schema is part of the record

```
<menu id="file" value="File">
  <popup>
    <menuitem value="New" onclick="CreateNewDoc()" />
    <menuitem value="Open" onclick="OpenDoc()" />
    <menuitem value="Close" onclick="CloseDoc()" />
  </popup>
</menu>
```

HTML

- A more relaxed version of XML for web pages

```
<!DOCTYPE html>
<html>
  <head>
    <title>My beautiful web page!</title>
  </head>
  <body>
    Here is my content.
  </body>
</html>
```

Tags & Elements

- Tags begin with < and end with >
- Usually occur in pairs

   Wow! creates a new element in the structure
- Elements can be nested

  This is a really interesting paragraph.
- Some tags never occur in pairs

  Usual to use trailing slash, but not necessary

Attributes

- HTML elements can be assigned attributes by including property/value pairs in the opening tag

```
<tagname property="value"></tagname>
```

- E.g., a link can be given an href attribute, whose value specifies the URL for that link

```
<a href="http://d3js.org/">The D3 website</a>
```

Classes and IDs

- Attributes that can be referenced later to identify specific pieces of content

```
<p class="awesome">Awe-inspiring paragraph</p>
```

- Elements can be assigned to multiple classes

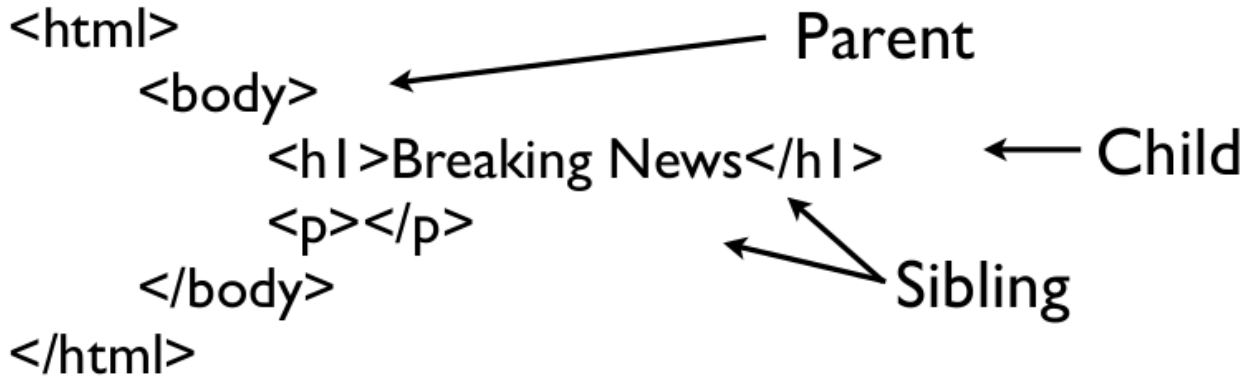
```
<p class="uplifting awesome">Awe-inspiring paragraph</p>
```

- IDs are similar, but only one ID per element, and each ID value only once per page

```
<div id="content">
        <div id="button"></div>
</div>
```

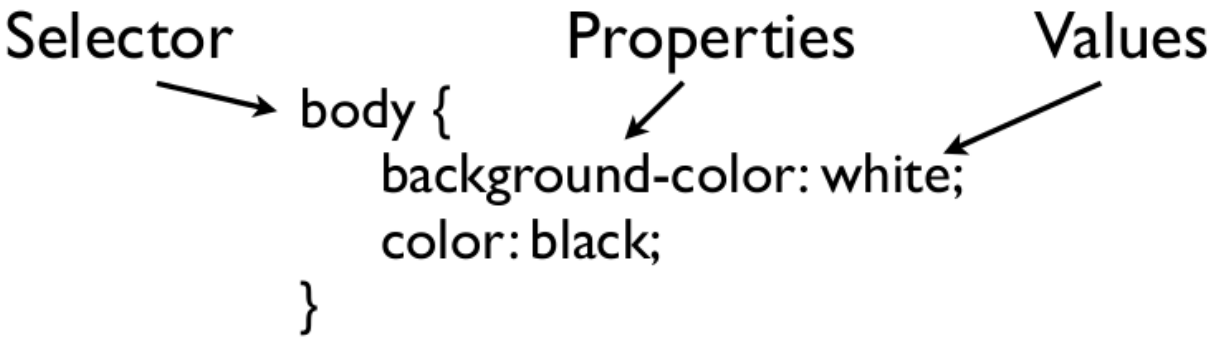
DOM

Document Object Model: the hierarchical structure of HTML



CSS

- Style the visual presentation of DOM elements



- Selectors come in different flavors

  Type selectors: h1, em, div, ...

  Class selectors: .caption, .label, .axis.x, ...

  ID selectors: #header, #nav, #button, ...

Referencing CSS

- Link to external CSS style sheets

```
<html>
  <head>
    <link rel="stylesheet" href="style.css">
  </head>
  <body>
    <p>Would you say I have style?</p>
  </body>
</html>
```

Web Scraping

Web Scraping[2]

"Data scraping is a technique in which a computer program extracts data from human-readable output coming from another program.
Web pages are built using text-based mark-up languages (HTML and XHTML), and frequently contain a wealth of useful data in text form. However, most web pages are designed for human end-users and not for ease of automated use. Because of this, tool kits that scrape web content were created. A web scraper is an API to extract data from a web site."

[2]Wikipedia

Inspecting Web Pages

- Chrome: DevTools
- Firefox: Developer Tools, Firebug
- Safari and Internet Explorer

Chrome Developer Tools Demo

Web Scraping in Python

- HTTP Requests: urllib2, requests
- HTML Parsing: lxml, beautifulsoup, pattern
- Crawling: Scrapy
- Controlling Browsers: Selenium/WebDriver
- Headless Browsers: PhantomJS

CSS Selectors[3]

- '*' : Selects all elements
- 'p' : Select all p tags
- '#myid' : Select element with id=myid
- '.myclass' : Select elements with class=myclass
- 'p #myid .myclass' : Union of all the selections
- 'div code': Find all code tags inside a div
- 'li > ul': Select all ul inside li (first level only)
- 'strong + em': Select all em that is immediately preceded by strong
- 'prev siblings': Selects all sibling elements that follow after the "prev" element, have the same parent, and match the filtering "siblings" selector.

[3] http://codylindley.com/jqueryselectors/

CSS Selectors[4]

- 'li[class]' : all li with attribute class
- [a="b"]: All elements with attribute with name 'a' with value 'b'
- 'li:first' : First element of li
- 'li:first-child' : First child of li
- 'li:even' : Even elements of li
- ':text' : All text boxes

[4] http://codylindley.com/jqueryselectors/

Crawling and Spiders

wget -mk -w 20 http://www.example.com/

- -m : Mirror a website
- -k : convert urls to point to local files
- -w : Delay between requests. 20 = 20 seconds, 20m => 20 minutes and so on
- -r : recursive download
- -p : download all files that are necessary to properly display a given HTML page.
- -c : continue a incomplete download
- –tries : Number of retries
- –reject, -A : File types to reject and accept

Scrapy[5]

```
from scrapy import Spider, Item, Field

class Post(Item):
    title = Field()

class BlogSpider(Spider):
    name, start_urls = 'blogspider',
                ['http://blog.scrapinghub.com']

    def parse(self, response):
        return [Post(title=e.extract())
                for e in response.css("h2 a::text")]
```

[5] http://scrapy.org/

Web Driver

- Originally a tool for automating testing of web applications
- Now a W3C standard (`http://w3c.github.io/webdriver/webdriver-spec.html`)
- Interface to Selenimum and WebDriver in Python

Selenium WebDriver[6]

```
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

driver = webdriver.Firefox()
driver.get("http://www.python.org")
elem = driver.find_element_by_name("q")
elem.send_keys("pycon")
elem.send_keys(Keys.RETURN)
```

[6] https://selenium-python.readthedocs.org/

PhantomJS[7]

- Scriptable Headless WebKit
- Automating web related workflow
- Use Cases:
  - Headless web testing
  - Page automation.
  - Screen capture.
  - Network monitoring.

[7] `https://github.com/ariya/phantomjs`

CasperJS[8]

- Navigation scripting and testing utility
- Built for PhantomJS
- Easy to define full navigation scenarios
- Syntactic sugar to make life very easy

[8] `http://casperjs.org/`

IMDB Web Scraping Demo

Facebook Web Scraping Demo

Summary

Major Concepts:

- Data Collection and Scraping
- Tools and Techniques

Slide Material References

- Slides from Harvard CS 109 (2013)
- Slides by Jeff Leek
- Also see slide footnotes