DATA SCIENCE
GETTING DATA

I.   **GETTING DATA**

II.   **REGEX / REQUESTS**

III. **API / WRAPPERS**

I. GETTING DATA

Data lives all over the internet

The question is whether or not
the author of the data makes it
easy for us to grab it.

II. REGEX / REQUESTS

REGular EXpressions

are how we capture patterns in text

AAPL

🔍

Finance

Apple Inc.   (NASDAQ:AAPL)

Add to portfolio

Company
   Summary
   News
   Option chain
   Related companies
   Historical prices
   Financials
Markets
News

112.82

+1.04 (0.93%)

Real-time: 3:18PM EST
NASDAQ real-time data - Discla
Currency in USD

Compare: Enter ticker here

Range 111.97 - 113.49   Div/yield 0.47/1.67
               19.75   EPS   6.43
            12.16   Shares   5.86B
          .32M   Beta   0.92
          .48B   Inst. own   62%
        7.54

g+1   7.6k

Back
Forward
Reload

Save As…
Print…
Translate to English
View Page Source
View Page Info

☰ Buffer This Page

Inspect Element

es ☐ Nasdaq ☐ MSFT ☐ SNDK ☐ SSNNF   **more »**

Dow
Nas
Tech
AAP

🔍 📱 **Elements** Network Sources Timeline Profiles

```
<div class="g-section sfe-break-bo
<script>…</script>
▼<div class="g-section g-tpl-right-
    ▶<div class="g-unit g-first">…</d
    ▼<div class="g-unit">
        ▼<div id="market-data-div" class="id-market-data-div nwp g-floatfix">
            ▼<div id="price-panel" class="id-price-panel goog-inline-block">
                ▶<div>…</div>
                ▶<div>…</div>
                </div>
            ▶<div class="snap-panel-and-plusone">…</div>
            </div>
        <script>google.finance.renderMarketData();</script>
        ▶<div id="sharebox-data" itemscope="itemscope" itemtype="http://schema.org/Intangible/FinancialQuote">…</div>
        </div>
    </div>
```
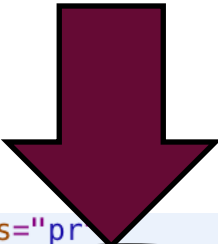
html body div #fjfe-real-body #fjfe-click-wrapper div #app #gf-viewc div div div div #market-data-div #price-panel div

Console Search Emulation Rendering

```
▼<span class="pr
   ▶<span id="ref_22144_l" class="unchanged">…</span>
 </span>
 ▶<div class="id-price-change nwp">…</div>
```

What regex can we use to capture this?

BEAUTIFULSOUP

Is a python based HTML parser.

WEB CRAWLERS

We just built one!

WEB CRAWLERS

We just built one!

But be careful….

III. APIS AND WRAPPERS

HTML Parsing   vs.   API

Must call using requests and
BeautifulSoup (imitate
human behavior)

Makes the call for us (the
author is "allowing us" to
access the data)

http://www.pythonforbeginners.com/api/list-of-python-apis

API (n):
Application Programming Interface

Easing access into a web based software

Examples of API's:

- Amazon (price data)
- Twitter (tweets)
- Facebook (social network)
- Sentiment Analysis

Examples of API's:

- Amazon (price data)
- Twitter (tweets)
- Facebook (social network)
- Sentiment Analysis

Mashape.com has an extensive collection

API   vs.   API wrapper

May still be a bit confusing
how to call the right page

Puts the API into a specific
programming language.
Gives us python functions.

http://www.pythonforbeginners.com/api/list-of-python-apis

Conclusion

Data is all over the web, but we must be polite and conscious of what data is available to us.