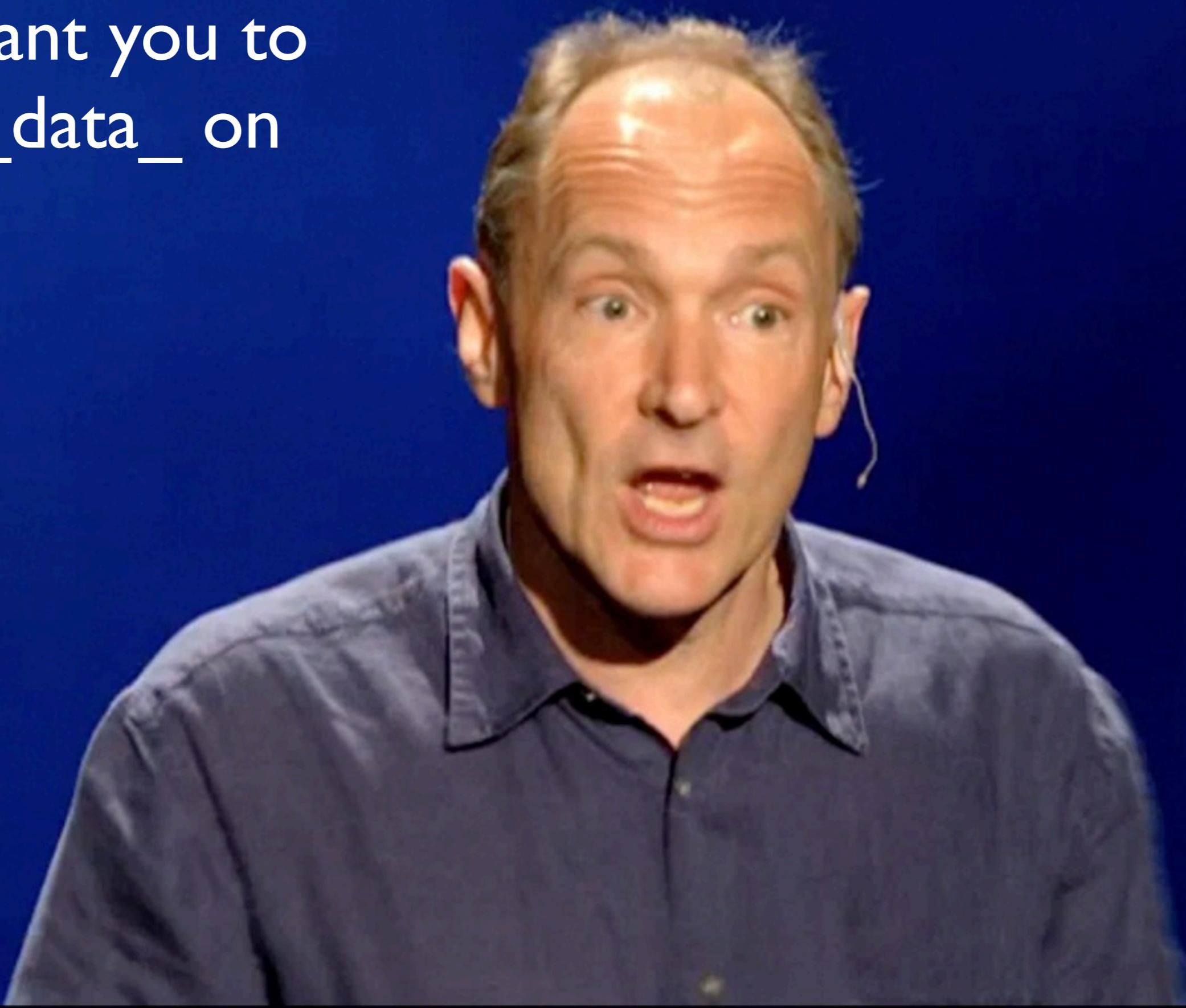


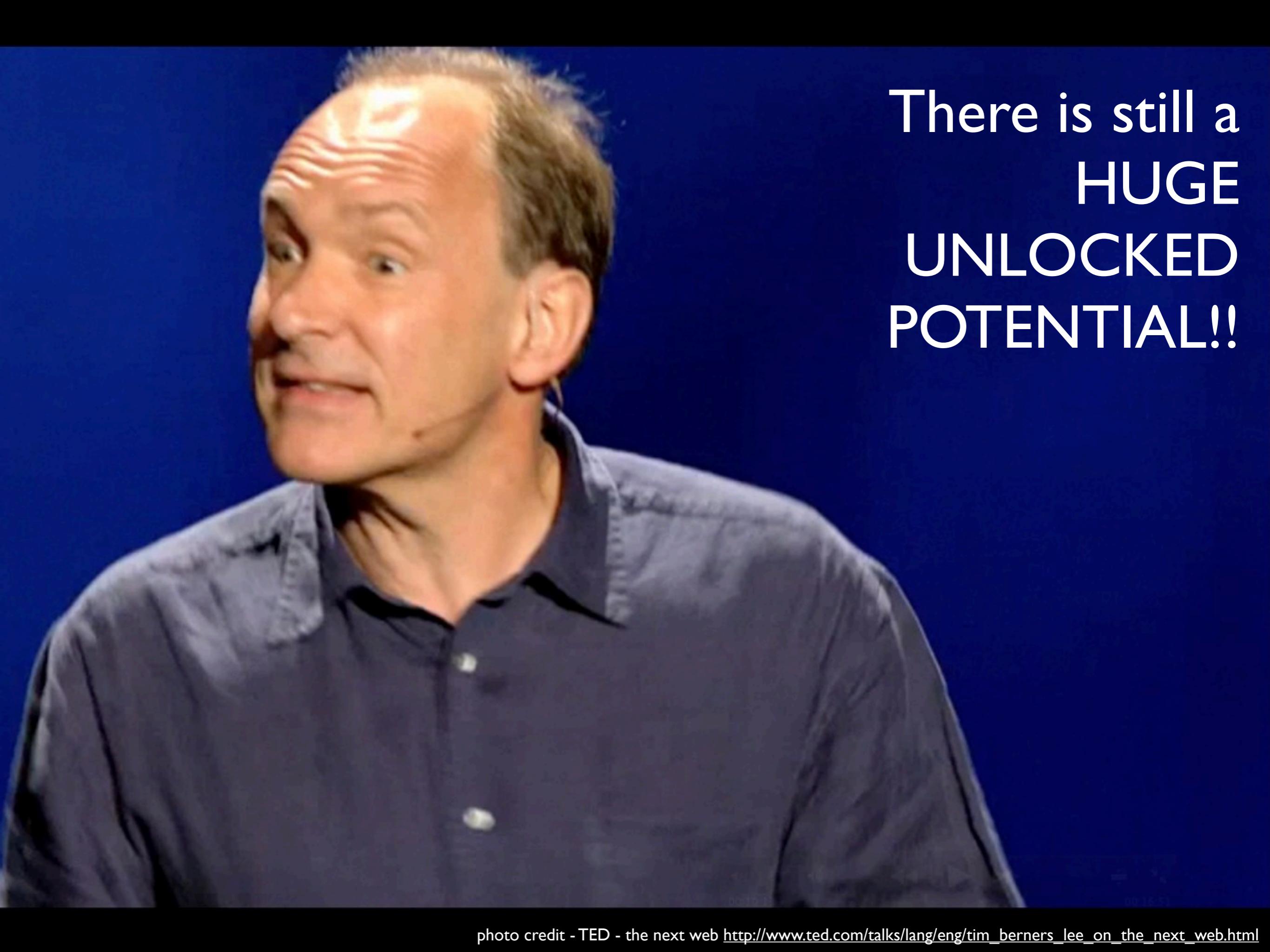
# Scraping with Python for Fun & Profit

@PyCon India 2010  
<http://in.pycon.org>

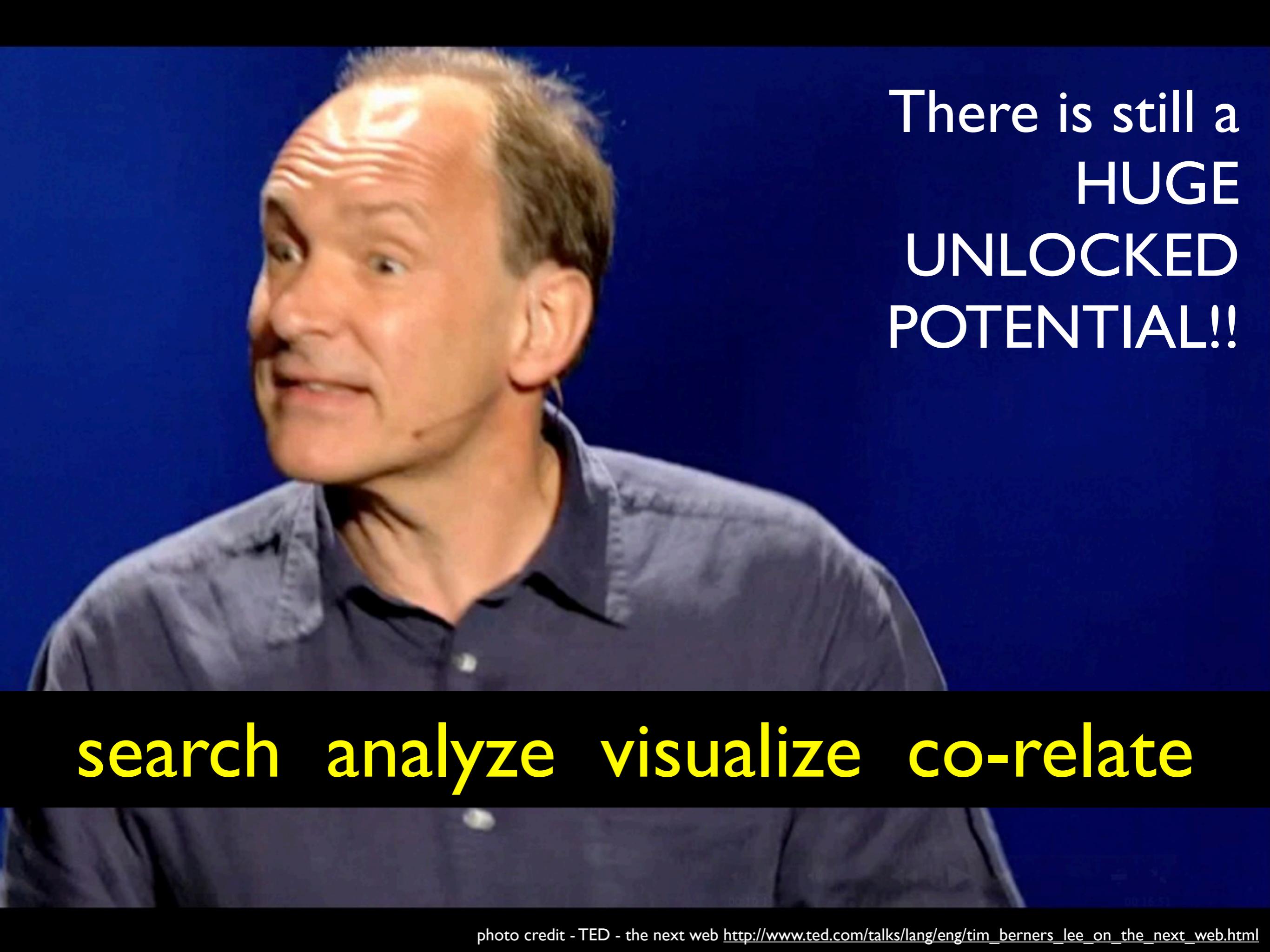
Abhishek Mishra  
[hello@ideamonk.com](mailto:hello@ideamonk.com)

Now... I want you to  
put your \_data\_ on  
the web!





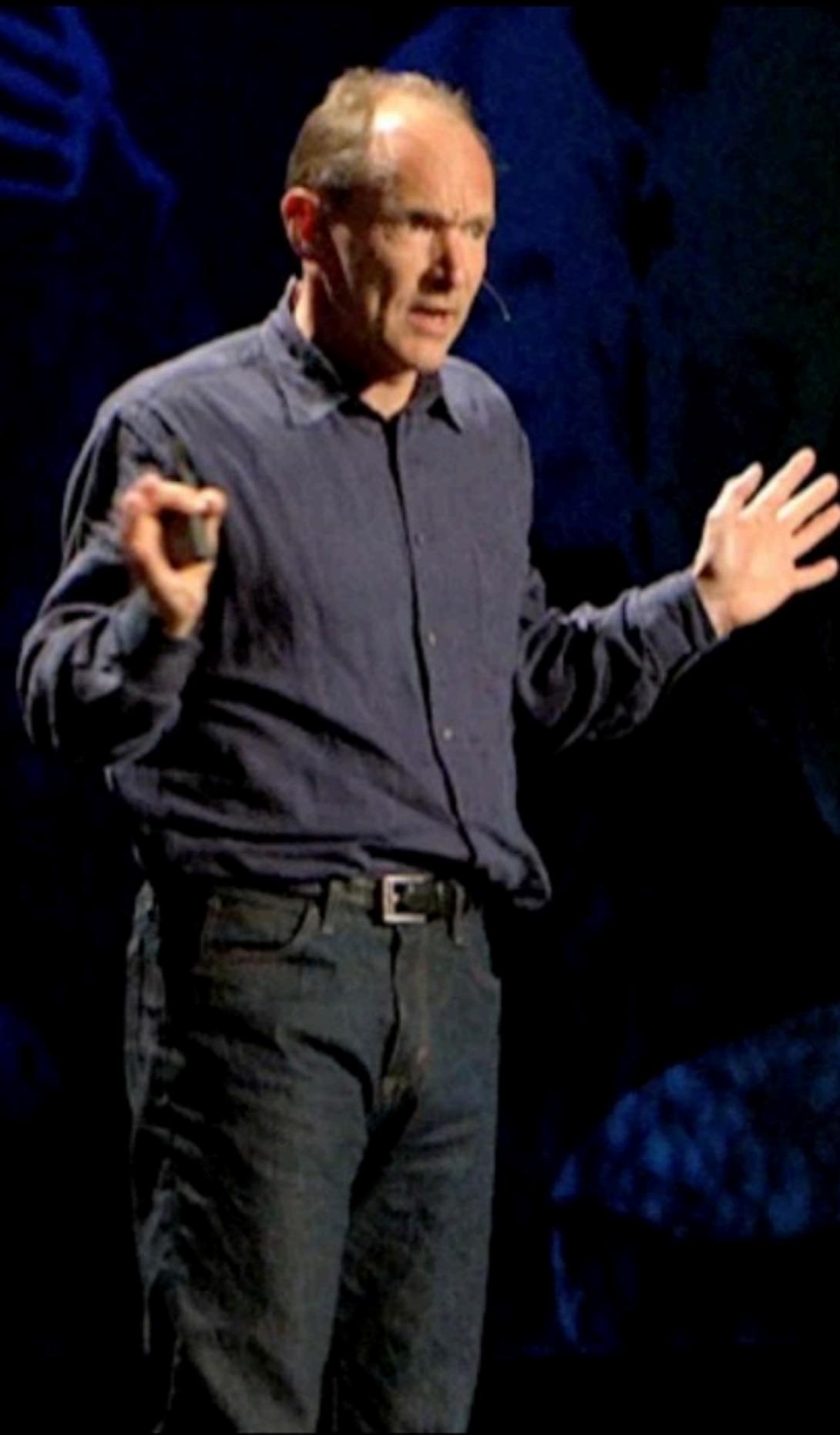
There is still a  
HUGE  
UNLOCKED  
POTENTIAL!!



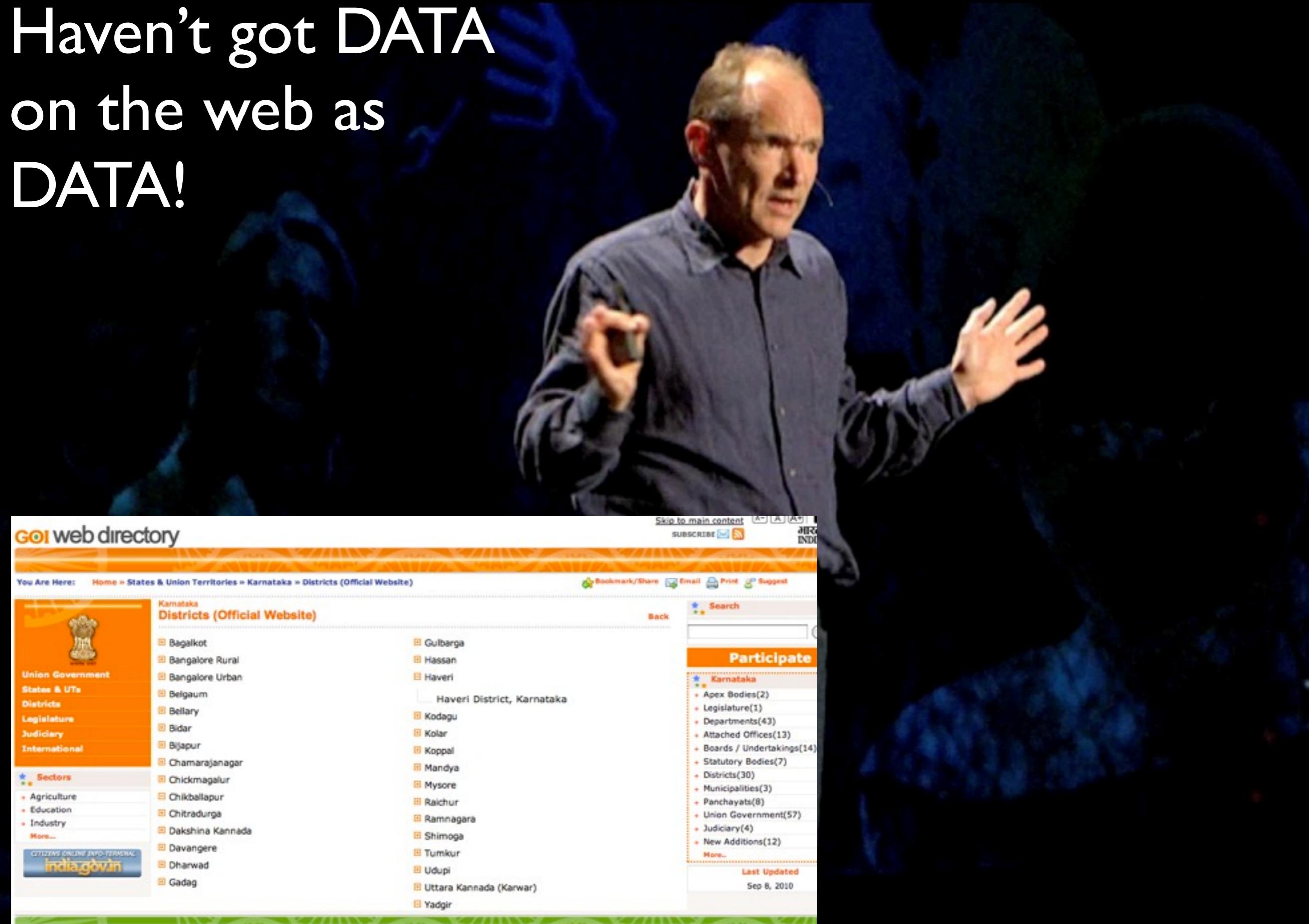
There is still a  
HUGE  
UNLOCKED  
POTENTIAL!!

search analyze visualize co-relate

Haven't got DATA  
on the web as  
DATA!



# Haven't got DATA on the web as DATA!



The image shows a man with short hair, wearing a dark blue button-down shirt, standing on a stage and gesturing with his hands while speaking. He appears to be giving a presentation or lecture. The background is dark with some blue lighting.

**GOI web directory**

You Are Here: Home > States & Union Territories > Karnataka > Districts (Official Website)

Karnataka Districts (Official Website)

Union Government  
States & UTs  
Districts  
Legislature  
Judiciary  
International

Sectors  
Agriculture  
Education  
Industry  
More...

CITIZENS ONLINE INFO-TERMINAL  
**india.gov.in**

Bagalkot Gulbarga  
Bangalore Rural Hassan  
Bangalore Urban Haveri  
Belgaum Haveri District, Karnataka  
Bellary Kodagu  
Bidar Kolar  
Bijapur Koppal  
Chamarajanagar Mandya  
Chickmagalur Mysore  
Chikballapur Raichur  
Chitradurga Ramnagara  
Dakshina Kannada Shimoga  
Davangere Tumkur  
Dharwad Udupi  
Gadag Uttara Kannada (Karwar)  
Yadgir

Bookmark/Share Email Print Suggest

Search

Participate

Karnataka  
Apex Bodies(2)  
Legislature(1)  
Departments(43)  
Attached Offices(13)  
Boards / Undertakings(14)  
Statutory Bodies(7)  
Districts(30)  
Municipalities(3)  
Panchayats(8)  
Union Government(57)  
Judiciary(4)  
New Additions(12)  
More...

Last Updated  
Sep 8, 2010

photo credit - TED - the next web [http://www.ted.com/talks/lang/eng/tim\\_berners\\_lee\\_on\\_the\\_next\\_web.html](http://www.ted.com/talks/lang/eng/tim_berners_lee_on_the_next_web.html)

# Haven't got DATA on the web as DATA!



A photograph of Tim Berners-Lee speaking on stage at a TED talk. He is wearing a dark shirt and has his hands clasped in front of him. The background is dark with some blue lighting.

648

649

650 <ul style="display:none" id="pKR009" class="bg\_link">

651

652

653 <a href="javascript:openChild('http://goidirectory.nic.in/sitecounter.php?id=1439','win2');" class="heading\_light\_url" title="http://chickmagalur.nic.in">Chickmagalur

654 District, Karnataka</a>

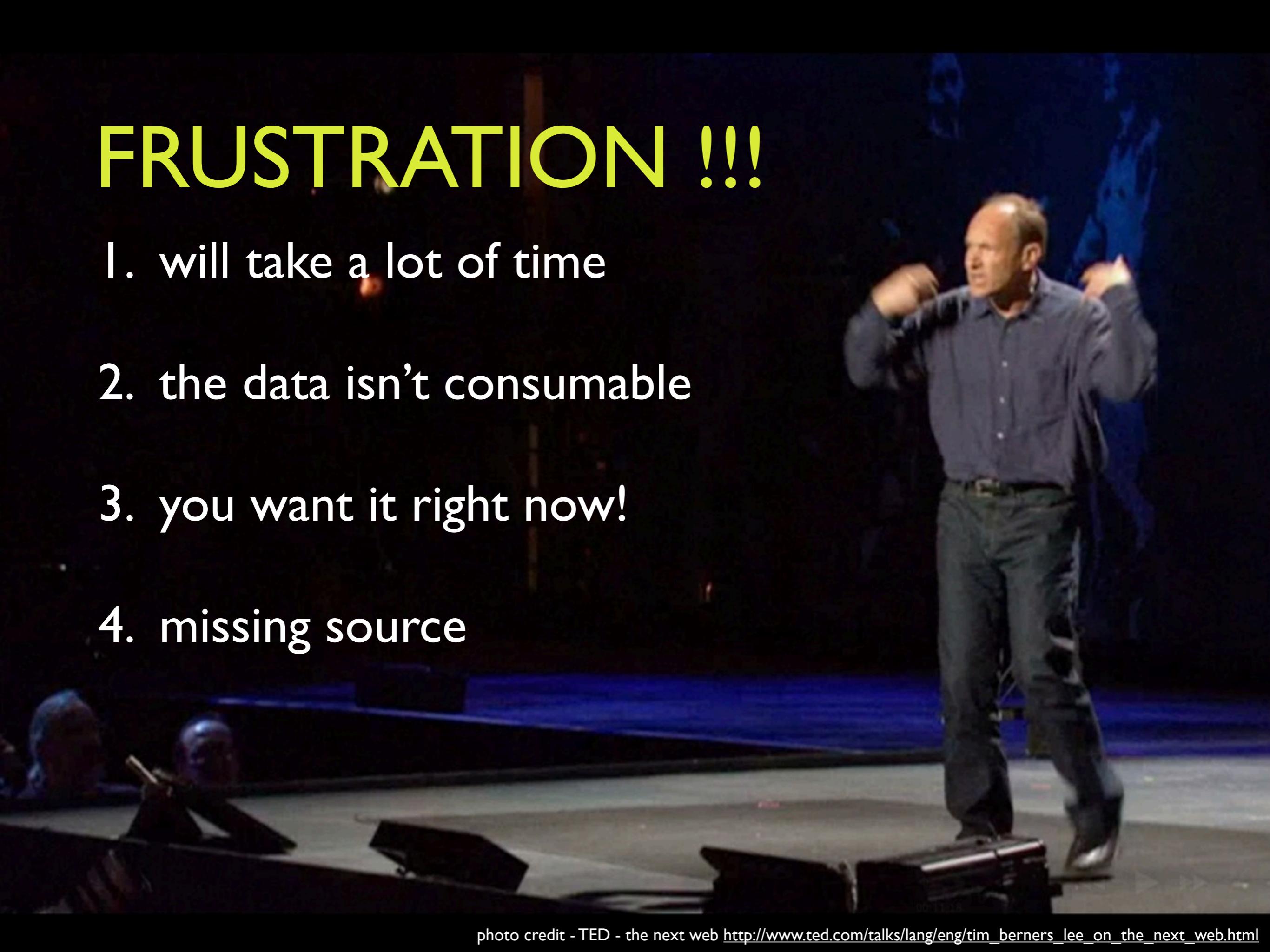
655 </li>

</ul>

The image shows a screenshot of a web page from the GOI web directory. The page displays a navigation menu on the left with links to various government sectors like Agriculture, Education, and Industry. The main content area shows a list item for Chickmagalur District, Karnataka, with a detailed URL in the href attribute of the anchor tag. The code is overlaid on the right side of the image, showing line numbers and the corresponding HTML code.

# FRUSTRATION !!!

1. will take a lot of time
2. the data isn't consumable
3. you want it right now!
4. missing source



# Solution ?





# Solution ?

just scrape it out!

# Solution ?

just **scrape** it out!

1. analyze request
2. find patterns
3. extract data

# Solution ?

just **scrape** it out!

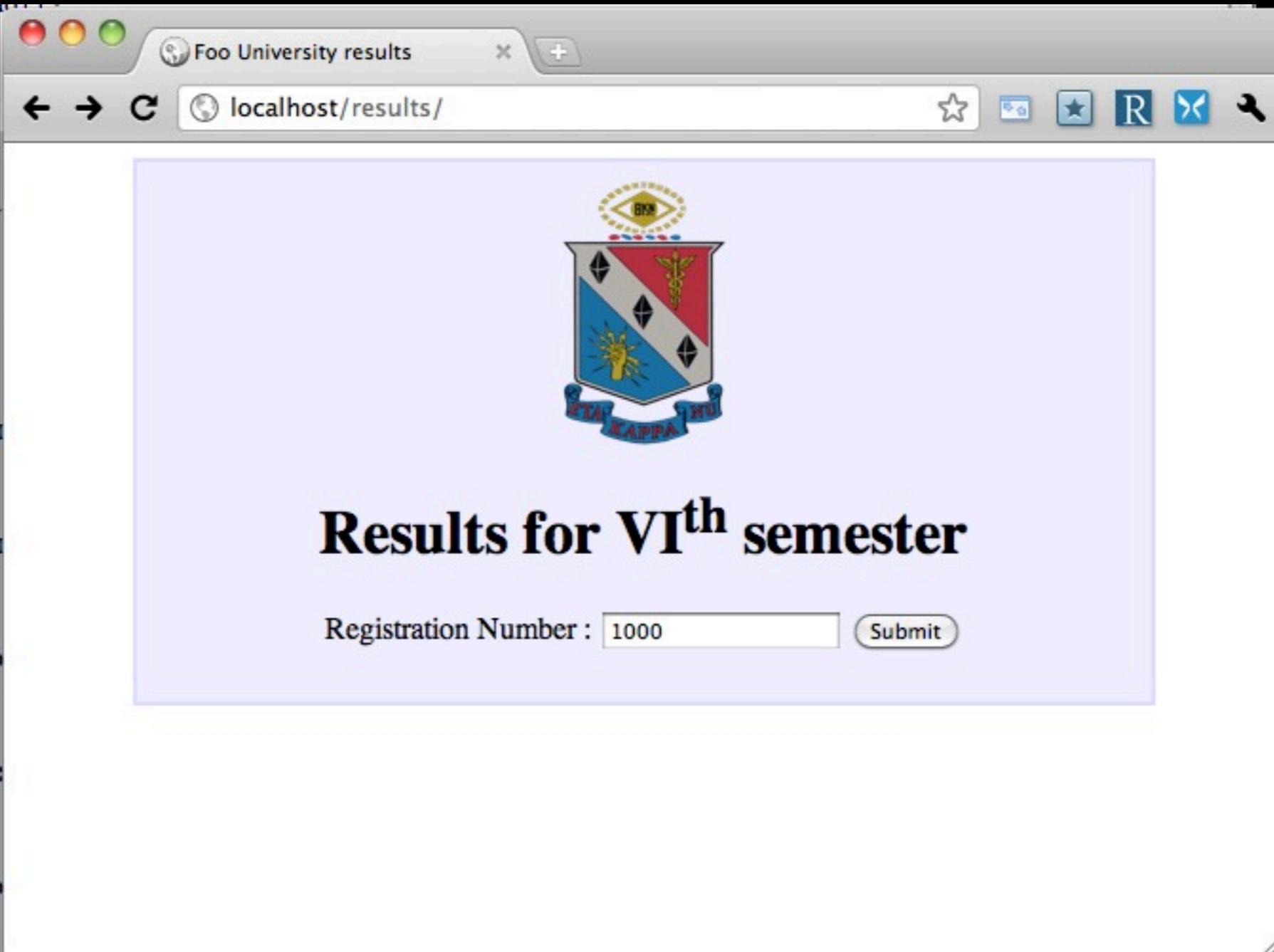
1. analyze request
2. find patterns
3. extract data

Use cases - **good bad ugly**

# Tools of Trade

- **urllib2**  
passing your data, getting back results
- **BeautifulSoup**  
parsing it out of the entangled web
- **Mechanize**  
programmatic web browsing
- **Scrapy**  
a web scraping framework

# Our objective



A screenshot of a web browser window titled "Foo University results". The URL in the address bar is "localhost/results/". The page content features a logo at the top center, which is a shield divided into four quadrants. The top left quadrant is blue with a white sun-like symbol, the top right is red with a caduceus, the bottom left is blue with a white diamond, and the bottom right is red with a white diamond. Above the shield is a small gold emblem with the letters "B.M.". Below the shield is a banner with the text "KAPPA" in white. The main heading "Results for VI<sup>th</sup> semester" is displayed prominently below the logo. A form field for "Registration Number" contains the value "1000", and a "Submit" button is to its right.

Results for VI<sup>th</sup> semester

Registration Number :

# Our objective

```
<body>
<div id="results">

<h1>Lohit Narayan
</h1>
<div>
Registration Number : 1234<br />
<table>
<tr>
<td>Object Oriented Programming</td>
<td> B+</td>
</tr>
<tr>
<td>Probability and Statistics</td>
<td> B+</td>
</tr>
<tr>
<td><b>CGPA</b></td>
<td><b>9.3</b></td>
</tr>
</table>
```

Name	Reg	CGPA
====	---	----
Lohit Narayan,	1234,	9.3
...	...	...

# urllib2

- Where to send?
- What to send?
- Answered by -  
firebug,  
chrome developer tools,  
a local debugging proxy

# urllib2

```
import urllib
import urllib2

url = “http://localhost/results/results.php”

data = { ‘regno’ : ‘1000’ }
data_encoded = urllib.urlencode(data)

request = urllib2.urlopen( url, data_encoded )
html = request.read()
```

# urllib2

```
import urllib  
import urllib2
```

```
url = “http://localhost/results/results.php”
```

```
data = { ‘regno’ : ‘1000’ }  
data_encoded = urllib.urlencode(data)
```

```
request = urllib2.urlopen( url, data_encoded )  
html = request.read()
```

```
'<html>\n<head>\n\t<title>Foo University results</title>\n\t<style type="text/css">\n\t\t#results {\n\t\tmargin:0px 60px;\n\t\tbackground: #efefff;\n\t\tpadding:10px;\n\t\tborder:2px solid #dd  
ddff;\n\t\ttext-align:center;\n\t}</style>\n</head>\n<body>\n\t<div id="results">\n\t<img sr  
c="shield.png" style="height:140px;"/>\n\t<h1>Rahim Kumar</h1>\n\t<div>\n\t\tRegistration Number :  
1000\n\t\t<br/>\n\t\t<table>\n\t\t\t<tr>\n\t\t\t\t<td>Humanities Elective II</td>\n\t\t\t\t<td>B-</td>\n\t\t\t\t<td>Humanities Elective III</td>\n\t\t\t\t<td>A</td>\n\t\t\t\t<td>Discrete Mathematics</td>\n\t\t\t\t<td>C-</td>\n\t\t\t\t<td>Environmental Science and Engineering</td>\n\t\t\t\t<td>B</td>\n\t\t\t\t<td>Data Structures</td>\n\t\t\t\t<td>D</td>\n\t\t\t\t<td>Science Elective I</td>\n\t\t\t\t<td>D-</td>\n\t\t\t\t<td>Object Oriented Programming</td>\n\t\t\t\t<td>B</td>\n\t\t\t\t<td>Probability and Statistics</td>\n\t\t\t\t<td>A-</td>\n\t\t\t\t<td>CGPA</td>\n\t\t<td>\n\t\t\t<b>9.1</b></td>\n\t\t</tr>\n\t\t</table>\n\t</div>\n\t</div>\n</body>\n</html>\n'
```

# urllib2

```
import urllib    # for urlencode
import urllib2   # main urllib2 module

url = “http://localhost/results/results.php”

data = { ‘regno’ : ‘1000’ }
# ^ data as an object

data_encoded = urllib.urlencode(data)
# a string, key=val pairs separated by ‘&’

request = urllib2.urlopen( url, data_encoded )
# ^ returns a network object

html = request.read()
# ^ reads out its contents to a string
```

# urllib2

```
import urllib
import urllib2

url = “http://localhost/results/results.php”
data = {}
scraped_results = []

for i in xrange(1000, 3000):
    data[‘regno’] = str(i)
    data_encoded = urllib.urlencode(data)
    request = urllib2.urlopen(url, data_encoded)
    html = request.read()
    scraped_results.append(html)
```

```
'<html>\n<head>\n<title>Foo University results</title>\n<style type="text/css">\n<#results {\n    margin: 0px 60px;\n    background: #efefff;\n    padding: 10px;\n    border: 2px solid #ddcff;\n    text-align: center;\n}>\n</head>\n<body>\n<div id="results">\n\n<h1>Rahim Kumar</h1>\n<div>\n    Registration Number : 1000\n    <br/>\n    <table>\n        <tr>\n            <td>Humanities Elective II</td>\n            <td>B-</td>\n            <td>Humanities Elective III</td>\n            <td>A</td>\n            <td>Discrete Mathematics</td>\n            <td>C-</td>\n            <td>Environmental Science and Engineering</td>\n            <td>B</td>\n            <td>Data Structures</td>\n            <td>D</td>\n            <td>Science Elective I</td>\n            <td>D-</td>\n            <td>Object Oriented Programming</td>\n            <td>B</td>\n            <td>Probability and Statistics</td>\n            <td>A-</td>\n        </tr>\n        <tr>\n            <td><b>CGPA</b></td>\n            <td>9.1</td>\n        </tr>\n    </table>\n</div>\n</div>\n</body>\n</html>'
```

Thats alright,  
But this is still meaningless.

```
'<html>\n<head>\n<title>Foo University results</title>\n<style type="text/css">\n<#results {\n    margin: 0px 60px;\n    background: #efefff;\n    padding: 10px;\n    border: 2px solid #ddcff;\n    text-align: center;\n}</style>\n</head>\n<body>\n<div id="results">\n\n<h1>Rahim Kumar</h1>\n<div>\n    Registration Number : 1000\n    <br/>\n    <table>\n        <tr>\n            <td>Humanities Elective II</td>\n            <td>B-</td>\n            <td>Humanities Elective III</td>\n            <td>A</td>\n            <td>Discrete Mathematics</td>\n            <td>C-</td>\n            <td>Environmental Science and Engineering</td>\n            <td>B</td>\n            <td>Data Structures</td>\n            <td>D</td>\n            <td>Science Elective I</td>\n            <td>D-</td>\n            <td>Object Oriented Programming</td>\n            <td>B</td>\n            <td>Probability and Statistics</td>\n            <td>A-</td>\n        </tr>\n        <tr>\n            <td><b>CGPA</b></td>\n            <td>9.1</td>\n        </tr>\n    </table>\n</div>\n</div>\n</body>\n</html>'
```

Thats alright,  
But this is still meaningless.

Enter BeautifulSoup

# BeautifulSoup



# BeautifulSoup

- Python HTML/XML Parser
- Won't choke you on bad markup
- auto-detects encoding
- easy tree traversal
- find all links whose url contains 'foo.com'

# BeautifulSoup

```
>>> import re
>>> from BeautifulSoup import BeautifulSoup

>>> html = """<a href="http://foo.com/foobar">link1</a> <a href="http://foo.com/beep">link2</a> <a href="http://foo.co.in/beep">link3</a>"""

>>> soup = BeautifulSoup(html)
>>> links = soup.findAll('a', {'href': re.compile('.*foo.com.*')})
>>> links
[<a href="http://foo.com/foobar">link1</a>, <a href="http://foo.com/beep">link2</a>]

>>> links[0]['href']
u'http://foo.com/foobar'

>>> links[0].contents
[u'link1']
```

# BeautifulSoup

```
>>> html = """<html>\n    <head>\n        <title>\n            Foo University results\n        </title>\n        <style type="text/css">\n            #results {\n                margin:0px 60px;\n                background: #efefff;\n                padding:10px;\n                border:2px solid #ddcff;\n                text-align:center;\n            }\n        </style>\n    </head>\n    <body>\n        <div id="results">\n            \n                Bipin Narayan\n            </h1>\n            <div>\n                Registration Number : 1000\n                <br />\n                <table>\n                    <tr>\n                        <td>\n                            Humanities Elective II\n                        </td>\n                        <td>\n                            C-\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Discrete Mathematics\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Environmental Science and Engineering\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Science Elective\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Object Oriented Programming\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Probability and Statistics\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            CGPA\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                </table>\n            </div>\n        </body>\n    </html>""""
```

# BeautifulSoup

```
>>> html = """<html>\n    <head>\n        <title>\n            Foo University results\n        </title>\n        <style type="text/css">\n            #results {\n                margin:0px 60px;\n                background: #efefff;\n                padding:10px;\n                border:2px solid #ddcff;\n                text-align:center;\n            }\n        </style>\n    </head>\n    <body>\n        <div id="results">\n            \n                Bipin Narayan\n            </h1>\n            <div>\n                Registration Number : 1000\n                <br />\n                <table>\n                    <tr>\n                        <td>\n                            Humanities Elective II\n                        </td>\n                        <td>\n                            C-\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Discrete Mathematics\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Environmental Science and Engineering\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Science Elective\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Object Oriented Programming\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Probability and Statistics\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            CGPA\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                </table>\n            </div>\n        </body>\n    </html>""""
```

```
>>> soup.h1
<h1>Bipin Narang</h1>
>>> soup.h1.contents
[u'Bipin Narang']
>>> soup.h1.contents[0]
u'Bipin Narang'
```

# BeautifulSoup

```
>>> html = """<html>\n    <head>\n        <title>\n            Foo University results\n        </title>\n        <style type="text/css">\n            #results {\n                margin:0px 60px;\n                background: #efefff;\n                padding:10px;\n                border:2px solid #ddcff;\n                text-align:center;\n            }\n        </style>\n    </head>\n    <body>\n        <div id="results">\n            \n                Bipin Narayan\n            </h1>\n            <div>\n                Registration Number : 1000\n                <br />\n                <table>\n                    <tr>\n                        <td>\n                            Humanities Elective II\n                        </td>\n                        <td>\n                            C-\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Discrete Mathematics\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Environmental Science and Engineering\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            8\n                        </td>\n                    </tr>\n                </table>\n                <tr>\n                    <td>\n                        Object Oriented Programming\n                    </td>\n                    <td>\n                        Probability and Statistics\n                    </td>\n                    <td>\n                        CGPA\n                    </td>\n                </tr>\n            </div>\n        </body>\n    </html>"""
```

```
>>> soup.div.div.contents[0]
u'\n\t\tRegistration Number : 1000\t\t'
>>> soup.div.div.contents[0].split(':')
[u'\n\t\tRegistration Number ', u' 1000\t\t']
>>> soup.div.div.contents[0].split(':')[1].strip()
u'1000'
```

# BeautifulSoup

```
>>> html = """<html>\n    <head>\n        <title>\n            Foo University results\n        </title>\n        <style type="text/css">\n            #results {\n                margin:0px 60px;\n                background: #efefff;\n                padding:10px;\n                border:2px solid #ddcff;\n                text-align:center;\n            }\n        </style>\n    </head>\n    <body>\n        <div id="results">\n            \n                Bipin Narayan\n            </h1>\n            <div>\n                Registration Number : 1000\n                <br />\n                <table>\n                    <tr>\n                        <td>\n                            Humanities Elective II\n                        </td>\n                        <td>\n                            C-\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            C\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <td>\n                            B-\n                        </td>\n                        <td>\n                            B+\n                        </td>\n                        <td>\n                            A\n                        </td>\n                        <b>\n                            8\n                        </b>\n                    </tr>\n                    <tr>\n                        <td>\n                            Humanities Elective III\n                        </td>\n                        <td>\n                            Discrete Mathematics\n                        </td>\n                        <td>\n                            Environmental Science and Engineering\n                        </td>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            II\n                        </td>\n                        <td>\n                            III\n                        </td>\n                        <td>\n                            IV\n                        </td>\n                        <td>\n                            V\n                        </td>\n                        <td>\n                            VI\n                        </td>\n                        <td>\n                            VII\n                        </td>\n                        <td>\n                            VIII\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Discrete Mathematics\n                        </td>\n                        <td>\n                            Environmental Science and Engineering\n                        </td>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            II\n                        </td>\n                        <td>\n                            III\n                        </td>\n                        <td>\n                            IV\n                        </td>\n                        <td>\n                            V\n                        </td>\n                        <td>\n                            VI\n                        </td>\n                        <td>\n                            VII\n                        </td>\n                        <td>\n                            VIII\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Environmental Science and Engineering\n                        </td>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            II\n                        </td>\n                        <td>\n                            III\n                        </td>\n                        <td>\n                            IV\n                        </td>\n                        <td>\n                            V\n                        </td>\n                        <td>\n                            VI\n                        </td>\n                        <td>\n                            VII\n                        </td>\n                        <td>\n                            VIII\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            Data Structures\n                        </td>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            II\n                        </td>\n                        <td>\n                            III\n                        </td>\n                        <td>\n                            IV\n                        </td>\n                        <td>\n                            V\n                        </td>\n                        <td>\n                            VI\n                        </td>\n                        <td>\n                            VII\n                        </td>\n                        <td>\n                            VIII\n                        </td>\n                    </tr>\n                    <tr>\n                        <td>\n                            I\n                        </td>\n                        <td>\n                            II\n                        </td>\n                        <td>\n                            III\n                        </td>\n                        <td>\n                            IV\n                        </td>\n                        <td>\n                            V\n                        </td>\n                        <td>\n                            VI\n                        </td>\n                        <td>\n                            VII\n                        </td>\n                        <td>\n                            VIII\n                        </td>\n                    </tr>\n                </table>\n                <td>\n                    \n            </div>\n        </body>\n    </html>"""
```

```
>>> soup.find('b')\n<b>CGPA</b>\n>>> soup.find('b').findNext('b')\n<b>8</b>\n>>> soup.find('b').findNext('b').contents[0]\nu'8'
```

# urllib2 + BeautifulSoup

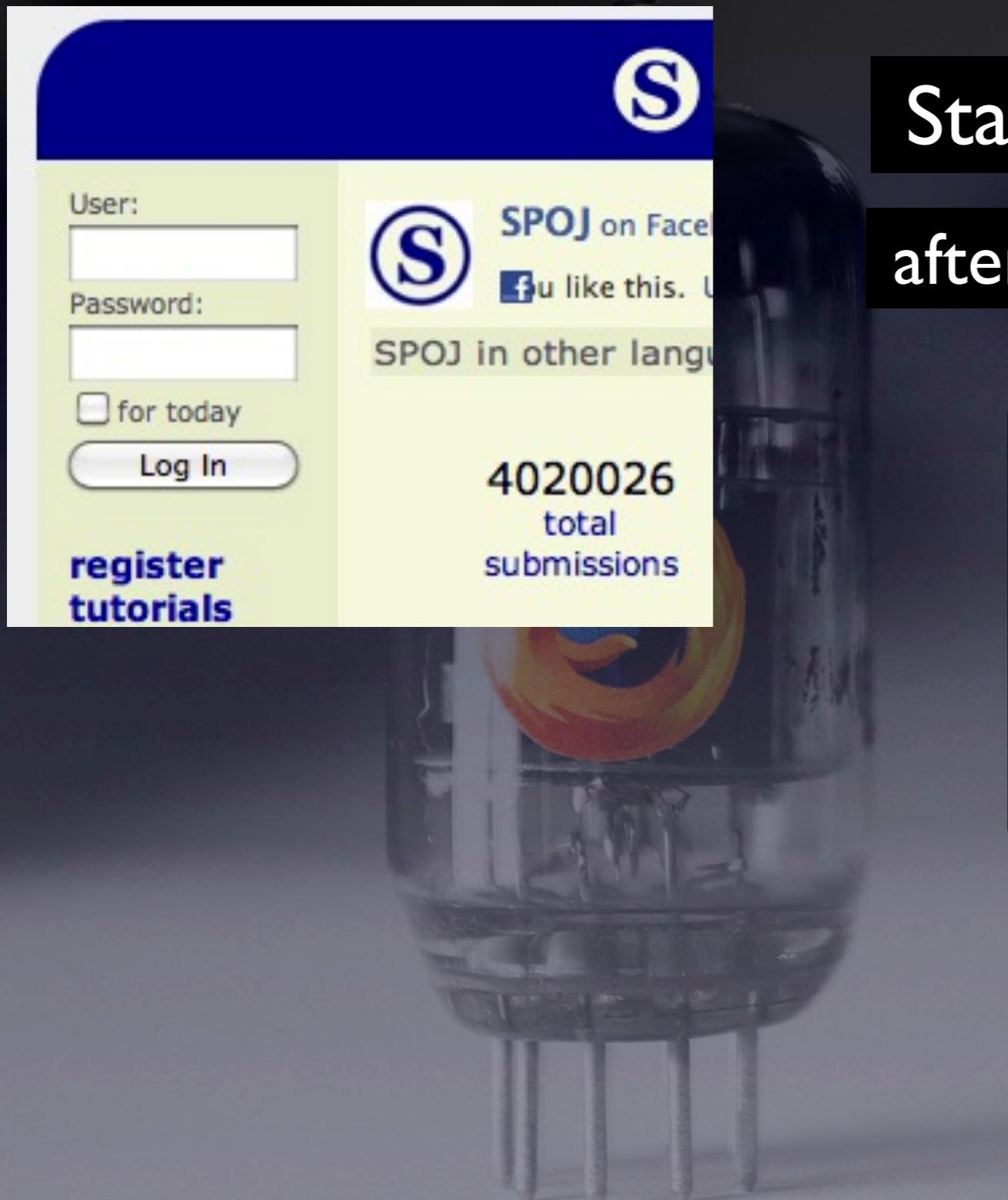
```
import urllib
import urllib2
from BeautifulSoup import BeautifulSoup

url = "http://localhost/results/results.php"
data={}
for i in xrange(1000,1100):
    data['regno'] = str(i)
    html = urllib2.urlopen( url, urllib.urlencode(data) ).read()
    soup = BeautifulSoup(html)
    name = soup.h1.contents[0]
    regno = soup.div.div.contents[0].split(':')[1].strip()
    # ^ though we already know current regno ;)
    cgpa = soup.find('b').findNext('b').contents[0]
    print "%s,%s,%s" % (regno, name, 'cgpa')
```

# Mechanize



# Mechanize



Stateful web browsing in Python

after Andy Lester's WWW:Mechanize

fill forms  
follow links  
handle cookies  
browser history

# Mechanize

```
>>> import mechanize  
>>> br = mechanize.Browser()  
>>> response = br.open("http://google.com")  
>>> br.title()  
'Google'  
>>> br.geturl()  
'http://www.google.co.in/'  
>>> print response.info()  
Date: Fri, 10 Sep 2010 04:06:57 GMT  
Expires: -1  
Cache-Control: private, max-age=0  
Content-Type: text/html; charset=ISO-8859-1  
Server: gws  
X-XSS-Protection: 1; mode=block  
Connection: close  
content-type: text/html; charset=ISO-8859-1  
>>> br.open("http://yahoo.com")  
>>> br.title()  
'Yahoo! India'  
>>> br.back()  
>>> br.title()  
'Google'
```

# Mechanize

## filling forms

```
>>> import mechanize  
>>> br = mechanize.Browser()  
>>> response = br.open("http://in.pycon.org/2010/account/  
login")  
>>> html = response.read()  
>>> html.find('href="/2010/user/ideamonk">Abhishek I')  
-1  
  
>>> br.select_form(name="login")  
>>> br.form["username"] = "ideamonk"  
>>> br.form["password"] = "*****"  
>>> response2 = br.submit()  
>>> html = response2.read()  
>>> html.find('href="/2010/user/ideamonk">Abhishek I')  
10186
```

# Mechanize

## filling forms

```
>>> import mechanize  
>>> br = mechanize.Browser()  
>>> response = br.open("http://in.pycon.org/2010/account/  
login")  
>>> html = response.read()  
>>> html.find('href="/2010/user/ideamonk">Abhishek I')  
-1  
  
>>> br.select_form(name="login")  
>>> br.form["username"] = "ideamonk"  
>>> br.form["password"] = "*****"  
>>> response2 = br.submit()  
>>> html = response2.read()  
>>> html.find('href="/2010/user/ideamonk">Abhishek I')  
10186
```

# Mechanize

# logging into gmail

```
<!-- login box -->
▼<div id="login">
  ►<script>...</script>
  ►<style type="text/css">...</style>
  ►<style type="text/css">...</style>
  ▼<form id="gaia_loginform" action="https://accounts.firefox.com/login?returnUrl=https%3A%2F%2Faccounts.firefox.com%2Flog&submit=Log+In" onsubmit="return(gaia_onLoginSubmit());">
    ▼<div id="gaia_loginbox">
      ▼<table class="form-noindent" cellspacing="0">
        ▼<tbody>
          ▼<tr>
            ▼<td valign="top" style="text-align: right; padding-right: 10px;">
              <input type="text" value="Email or phone number" name="username" style="width: 100%; height: 1.2em; border: 1px solid #ccc; border-radius: 3px; font-size: 1em; font-family: inherit; margin-bottom: 5px;" />
              <input type="password" value="Password" name="password" style="width: 100%; height: 1.2em; border: 1px solid #ccc; border-radius: 3px; font-size: 1em; font-family: inherit; margin-bottom: 5px;" />
              <input type="checkbox" checked="" value="Remember me" name="remember_me" style="width: 1.2em; height: 1.2em; border: 1px solid #ccc; border-radius: 3px; margin-bottom: 5px;" />
              <input type="button" value="Log In" style="width: 100%; height: 1.2em; background-color: #0070C0; color: white; border: none; border-radius: 3px; font-size: 1em; font-family: inherit; padding: 0 10px; margin-bottom: 5px;" />
              <small style="font-size: 0.8em; font-family: inherit; margin-bottom: 5px;">Forgot your password?
              <small style="font-size: 0.8em; font-family: inherit; margin-bottom: 5px;">Create an account
              <small style="font-size: 0.8em; font-family: inherit; margin-bottom: 5px;">Log in with SSO
            ▼<div style="margin-top: 10px; font-size: 0.8em; font-family: inherit; color: #0070C0; opacity: 0.8; transition: opacity 0.2s; cursor: pointer;">Don't have an account? Create one</div>
          ▼<div style="margin-top: 10px; font-size: 0.8em; font-family: inherit; color: #0070C0; opacity: 0.8; transition: opacity 0.2s; cursor: pointer;">Forgot your password? Reset it</div>
        ▼<div style="margin-top: 10px; font-size: 0.8em; font-family: inherit; color: #0070C0; opacity: 0.8; transition: opacity 0.2s; cursor: pointer;">Log in with SSO</div>
      ▼<div style="margin-top: 10px; font-size: 0.8em; font-family: inherit; color: #0070C0; opacity: 0.8; transition: opacity 0.2s; cursor: pointer;">Create an account</div>
    ▼<div style="margin-top: 10px; font-size: 0.8em; font-family: inherit; color: #0070C0; opacity: 0.8; transition: opacity 0.2s; cursor: pointer;">Forgot your password? Reset it</div>
  ▼<div style="margin-top: 10px; font-size: 0.8em; font-family: inherit; color: #0070C0; opacity: 0.8; transition: opacity 0.2s; cursor: pointer;">Log in with SSO</div>
</form>

```

# Mechanize

logging into gmail

```
▼<form id="gaia_loginform" action="https://  
"return(gaia_onLoginSubmit());">
```

the form has no name!  
cant **select\_form** it

# Mechanize

logging into gmail

```
>>> br.open("http://mail.google.com")
>>> login_form = br.forms().next()
>>> login_form["Email"] = "ideamonk@gmail.com"
>>> login_form["Passwd"] = "*****"
>>> response = br.open( login_form.click() )
>>> br.geturl()
'https://mail.google.com/mail/?shva=1'
>>> response2 = br.open("http://mail.google.com/mail/h/")
>>> br.title()
'Gmail - Inbox'
>>> html = response2.read()
```

# Mechanize

## reading my inbox

```
>>> soup = BeautifulSoup(html)
>>> trs = soup.findAll('tr', {'bgcolor': '#ffffff'})
>>> for td in trs:
...     print td.b
...
<b>madhukargunjan</b>
<b>NIHI</b>
<b>Chamindra</b>
<b>KR2 blr</b>
<b>Imran</b>
<b>Ratnadeep Debnath</b>
<b>yasir</b>
<b>Glenn</b>
<b>Joyent, Carly Guarcello</b>
<b>Zack</b>
<b>SahanaEden</b>
...
>>>
```

# Mechanize



Can scrape my gmail chats !?

# Scrapy



a **Framework** for web scraping

# Scrapy



uses **XPath**  
navigate through elements

# Scrapy

## XPath and Scrapy

```
<h1> My important chunk of text </h1>  
//h1/text()
```

```
<div id="description"> Foo bar beep &gt; 42 </div>  
//div[@id='description']
```

```
<div id='report'> <p>Text1</p> <p>Text2</p> </div>  
//div[@id='report']/p[2]/text()
```

- // select from the document, wherever they are
- / select from root node
- @ select attributes

scrapy.selector.XmlXPathSelector  
scrapy.selector.HtmlXPathSelector

<http://www.w3schools.com/xpath/default.asp>

# Scrapy



Interactive Scraping Shell!

\$ scrapy-ctl.py shell http://my/target/website

# Scrapy



Step I  
define a model to store Items

# Scrapy



Step II  
create your **Spider** to extract **Items**

# Scrapy



Step III

write a Pipeline to store them

# Scrapy

Lets write a Scrapy spider

# SpojBackup

- Spoj? Programming contest website
- 4021039 code submissions
- 81386 users
- web scraping == backup tool
- uses Mechanize
- <http://github.com/ideamonk/spojbackup>

# SpojBackup

- <https://www.spoj.pl/status/ideamonk/signedlist/>

```
*****
*   Signed document issued by Sphere Online Judge (http://www.spoj.pl)
*****
```

----- BEGIN OF DATA -----  
Submissions by [ideamonk] in contest [SPOJ]. DoI: 2010-09-10 12:18:26.

ID	DATE	PROBLEM	RESULT	TIME	MEM	LNG
3198919	2010-01-22 01:00:45	BOXES	AC	0.07	2544	C++
3023085	2009-11-29 14:54:36	ACODE	AC	0.46	5832	PYT
3023081	2009-11-29 14:54:15	ACODE	WA	0.68	5844	PYT
3023078	2009-11-29 14:53:38	ACODE	AC	0.47	5836	PYT
3023066	2009-11-29 14:50:22	ACODE	AC	0.37	5560	PYT
3023061	2009-11-29 14:50:00	ACODE	CE	0.00	0	PYT
3023030	2009-11-29 14:46:01	ACODE	AC	0.54	5556	PYT
3022963	2009-11-29 14:36:16	ACODE	RE	0.14	4124	PYT

- <https://www.spoj.pl/files/src/save/3198919>
- Code walkthrough

# Conclusions



# Conclusions

Python - large set of tools for scraping

Choose them wisely

Do not steal

# Conclusions

Python - large set of tools for scraping

Choose them wisely

Do not steal

## License and Site Access

IMDb grants you a limited license to access and make personal use of this site and not to download (other than page caching) or modify it, or any portion of it, except with express written consent of IMDb. This site or any portion of this site may not be reproduced, duplicated, copied, sold, resold, visited, or otherwise exploited for any commercial purpose without express written consent of IMDb. This license does not include any resale or commercial use of this site or its contents or any derivative use of this site or its contents.

**Robots and Screen Scraping:** You may not use data mining, robots, screen scraping, or similar data gathering and extraction tools on this site, except with our express written consent as noted below.

**Framing:** You may not frame or utilize framing techniques to enclose any trademark, logo, or other proprietary information (including images, text, page layout, or form) of IMDb without express written consent.

# Conclusions

Python - large set of tools for scraping

Choose them wisely

Do not steal

Use the cloud - [http://docs.picloud.com/advanced\\_examples.html](http://docs.picloud.com/advanced_examples.html)

Share your scrapers - <http://scraperwiki.com/>

Thanks to @amatix , flickr, you, etc

@PyCon India 2010  
<http://in.pycon.org>

Abhishek Mishra  
[hello@ideamonk.com](mailto:hello@ideamonk.com)