Python Library: lxml

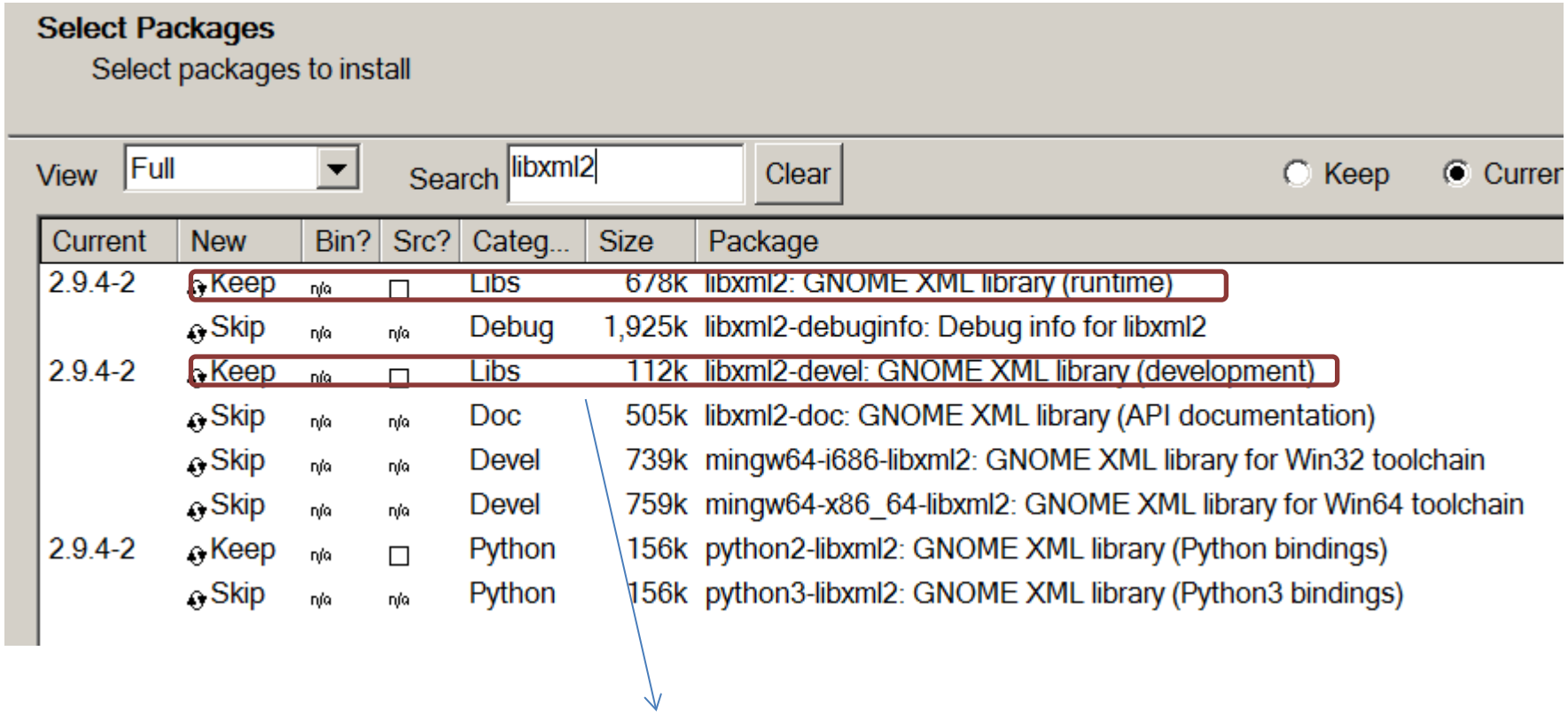
INF 55x

Wensheng Wu

Extraction using Python library lxml

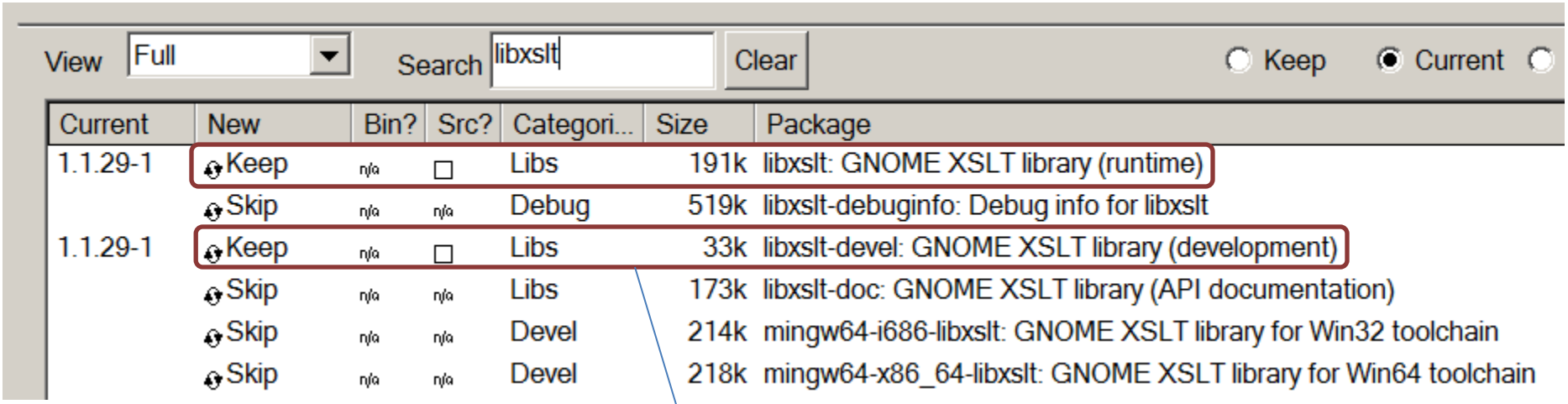
- lxml should be installed by default in your Amazon EC2 AMI Linux instance
- If not
  - sudo yum install libxml2-dev libxslt1-dev
  - sudo pip install lxml
- If you are using Cygwin
  - First install libxml2, libxslt (see next slide)
  - Then "pip install lxml"

Cygwin



Install libxml2 and libxml2-devel

Cygwin



Install libxslt and libxslt-devel

```
<bib>
  <cd>abc</cd>
  <book>
    <publisher>Addison-Wesley</publisher>
    <author>Serge Abiteboul</author>
    <author>
      <first-name>Rick</first-name>
      <last-name>Hull</last-name>
    </author>
    <author age="20">Victor Vianu</author>
    <title>Foundations of Databases</title>
    <year>1995</year>
    <price>38.8</price>
  </book>
  <book price="55">
    <publisher>Freeman</publisher>
    <author>Jeffrey D. Ullman</author>
    <title>Principles of Database and Knowledge Base Systems</title>
    <year>1998</year>
  </book>
  <book>
    <title>xyz</title>
    <author/>
  </book>
</bib>
```

Example

- from lxml import etree
- f = open('bibs.xml')
- tree = etree.parse(f)

- print(etree.tostring(tree, pretty_print=True))

Example

- for element in tree.xpath("//author"):
  print(etree.tostring(element))

```
<author>Serge Abiteboul</author>

<author><first-name>Rick</first-name><last-name>Hull</last-name></author>

<author age="20">Victor Vianu</author>

<author>Jeffrey D. Ullman</author>

<author/>
```

Example

- for element in tree.xpath("//author"):
  print element.tag, element.text

=>

author Serge Abiteboul
author None
author Victor Vianu
author Jeffrey D. Ullman
author None

Example

- for element in tree.xpath('//author[first-name="Rick"]'):

        print(etree.tostring(element))

=>

<author><first-name>Rick</first-name><last-name>Hull</last-name></author>

Helper function

- def printf(elements):

    for element in elements:

        print(etree.tostring(element))

- printf(tree.xpath('//author[first-name="Rick"]'))

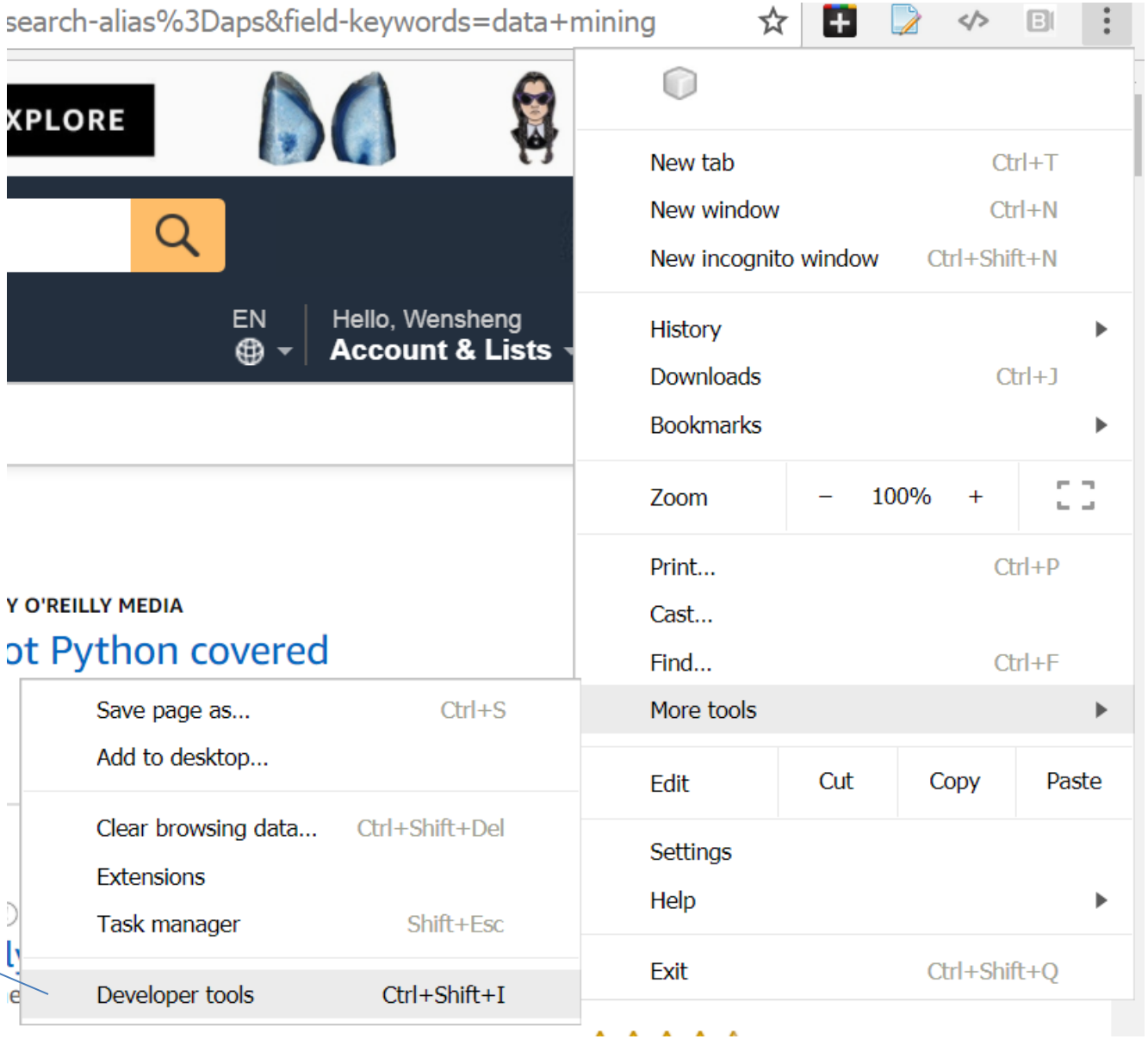
Work with HTML document

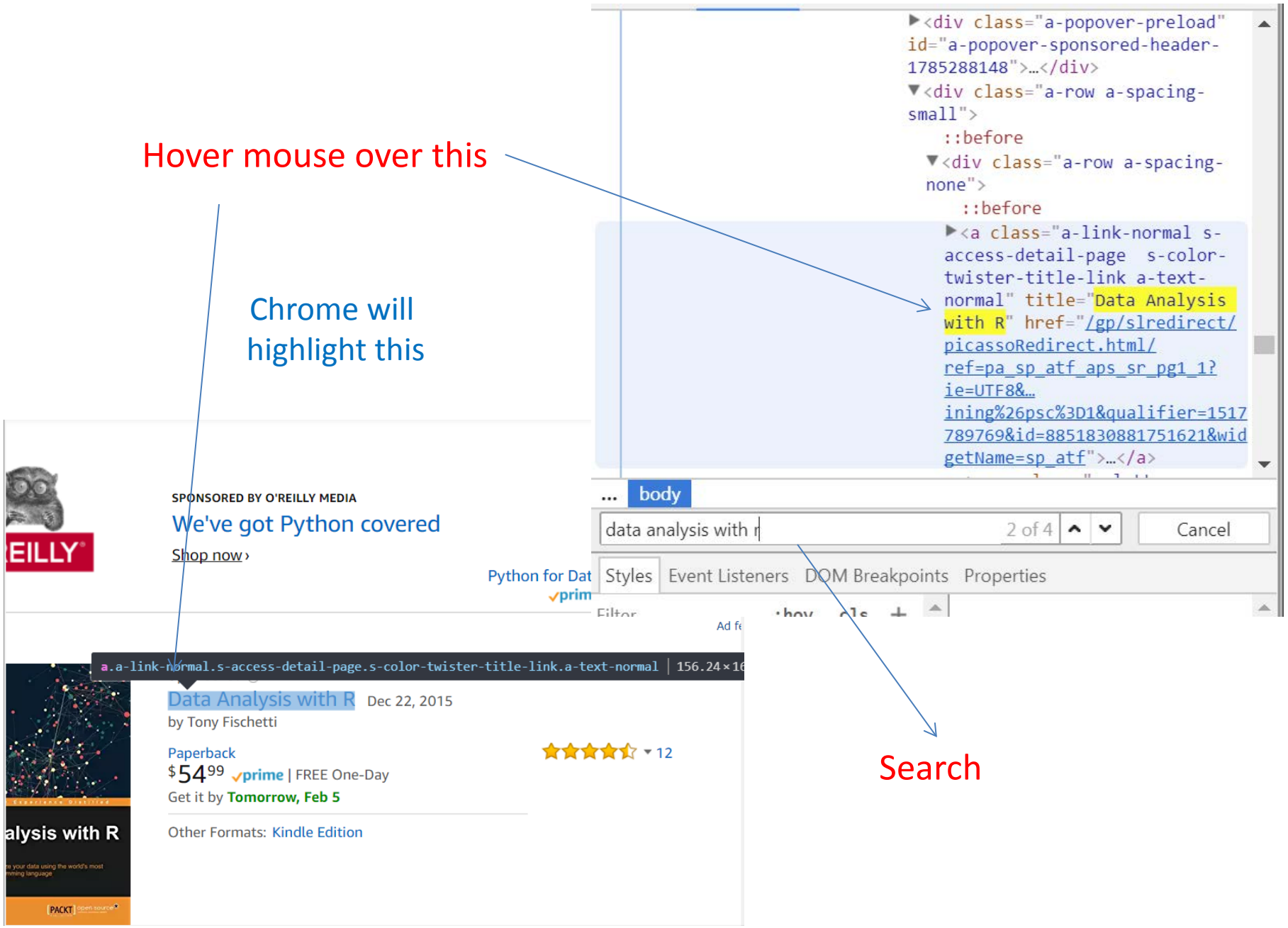
```
from lxml import html

myfile = open('Express.html')

data = myfile.read()
htree = html.fromstring(data)
```

Work with HTML document

```
print(etree.tostring(htree, pretty_print=True))


htree.xpath('//tbody/tr[1]/td[1]/text()')
htree.xpath('//tbody/tr[1]/td[2]/text()')
htree.xpath('//tbody/tr[1]/td[3]/text()')
```

Developer tools in Google Chrome




Hover mouse over this
Chrome will highlight this
…

::before

::before
…
body
data analysis with r
2 of 4
Cancel
Styles
Event Listeners
DOM Breakpoints
Properties
SPONSORED BY O'REILLY MEDIA
We've got Python covered
Shop now ›
a.a-link-normal.s-access-detail-page.s-color-twister-title-link.a-text-normal | 156.24×16
Data Analysis with R
Dec 22, 2015
by Tony Fischetti
Paperback
$54.99 prime | FREE One-Day
Get it by Tomorrow, Feb 5
12
Other Formats: Kindle Edition
Search