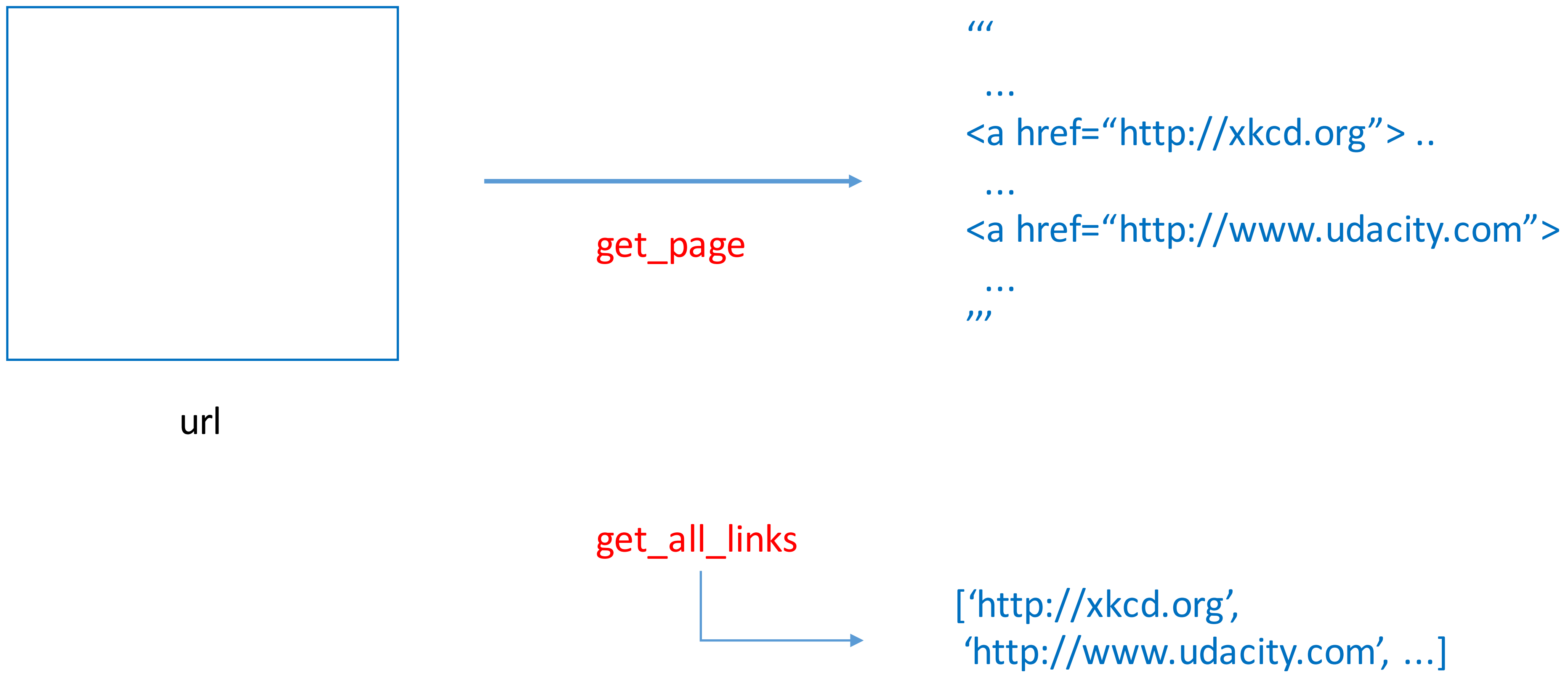
Intro to Computer Science

Local Laboratory

*** Udacity – Intro to Computer Science**

Collecting Links



url

get_page

```
"""
  ...
<a href="http://xkcd.org"> ..
  ...
<a href="http://www.udacity.com">
  ...
"""
```

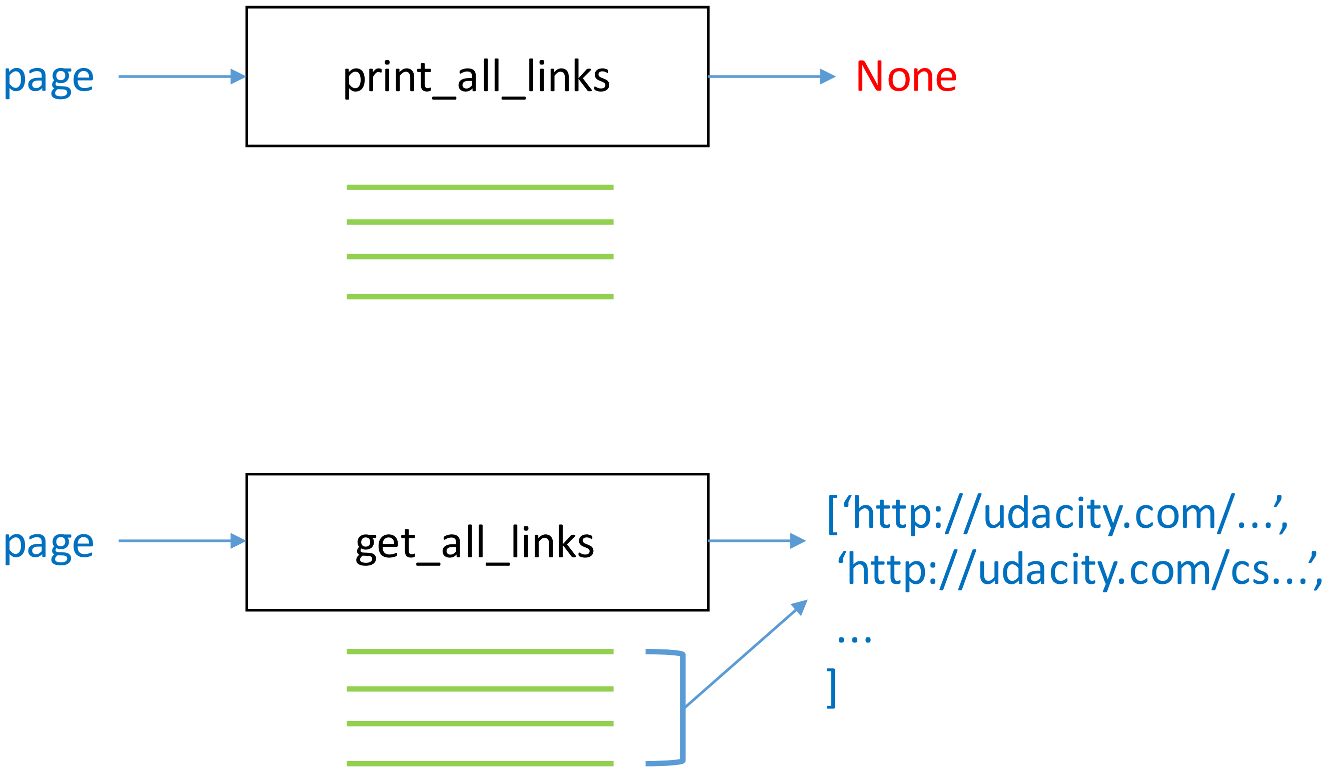
get_all_links

```
['http://xkcd.org',
 'http://www.udacity.com', ...]
```

Get All Links

```
def get_next_target(page):
    start_link = page.find('<a href=')
    if start_link == -1:
        return None, 0
    start_quote = page.find('"', start_link)
    end_quote = page.find('"', start_quote + 1)
    url = page[start_quote + 1:end_quote]
    return url, end_quote

def print_all_links(page):
    while True:
        url,endpos = get_next_target(page)
        if url:
            print(url)
            page = page[endpos:]
        else:
            break
```



page

print_all_links

None

page

get_all_links

[‘http://udacity.com/…’,
‘http://udacity.com/cs…’,
…
]

Links

code

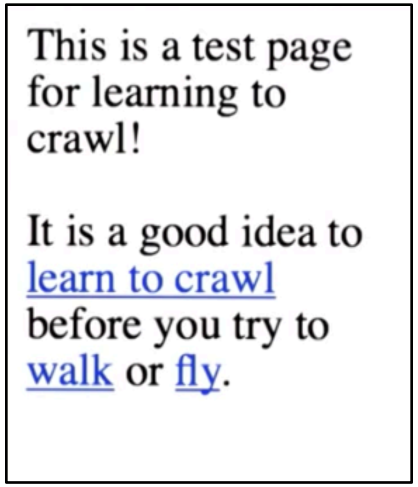
```
link = get_all_links(get_page('http://www.udacity.com/cs101x/index.html'))
print(link)
print(link[0])
```

result

```
['http://www.udacity.com/cs101x/crawling.html',
'http://www.udacity.com/cs101x/walking.html',
'http://www.udacity.com/cs101x/flying.html']

http://www.udacity.com/cs101x/crawling.html
```

../cs101/index.html'



This is a test page for learning to crawl!

It is a good idea to learn to crawl before you try to walk or fly.

Quiz: Starting Get All Links

```
def   get _all_links(page):
    links = □
    while True:
        url,endpos = get_next_target(page)
        if url:
            print(url)
            page = page[endpos:]
        else:
            break
```

Quiz: Updating Links

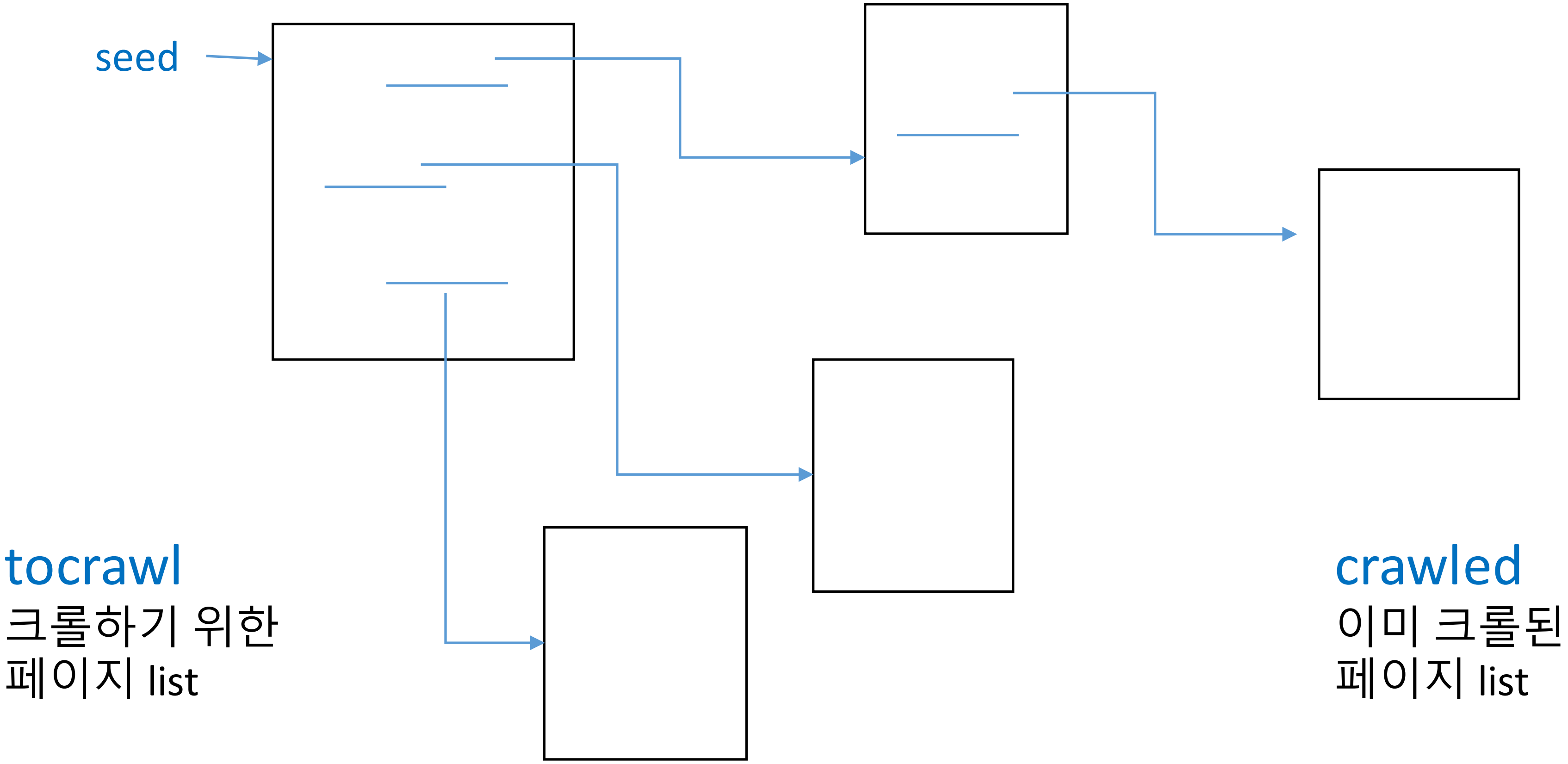
```
def    get_all_links(page):
    links = []
    while True:
        url,endpos = get_next_target(page)
        if url:

            page = page[endpos:]
        else:
            break
```

Quiz: Finishing Get All Links

```
def get_all_links(page):
    links = []
    while True:
        url,endpos = get_next_target(page)
        if url:
            links.append(url)
            page = page[endpos:]
        else:
            break
    return links
```

Finishing the Web Crawler



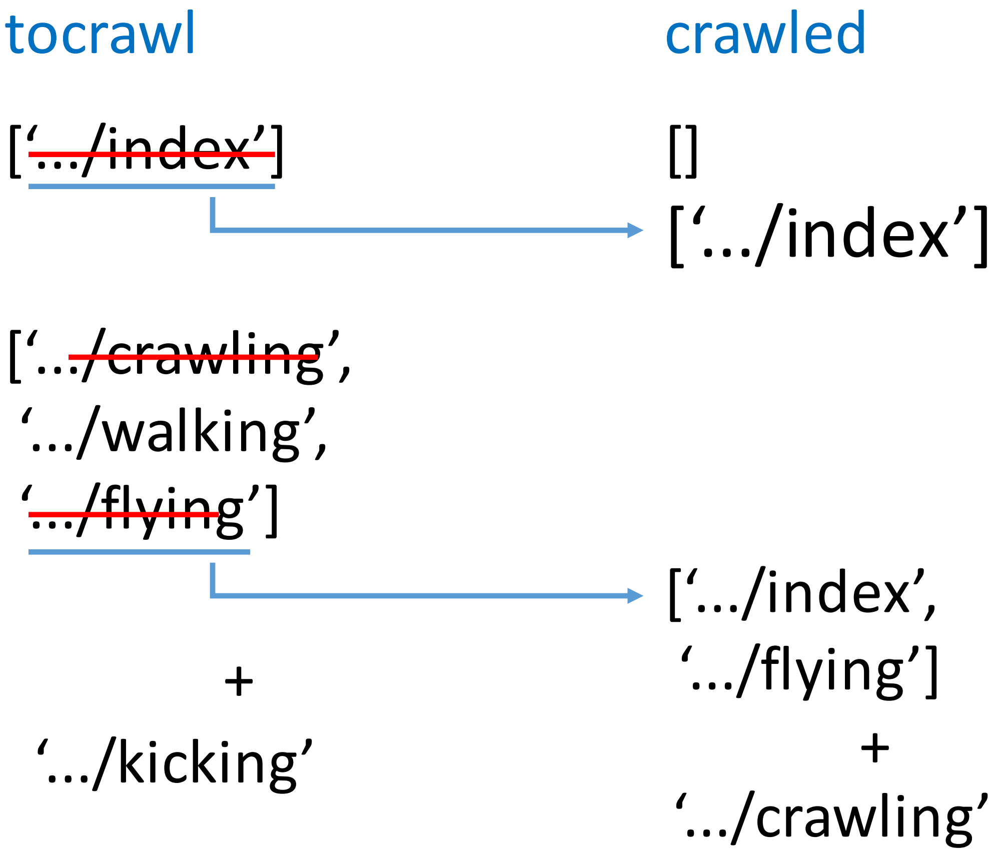
seed

tocrawl
크롤하기 위한
페이지 list

crawled
이미 크롤된
페이지 list

Finishing the Web Crawler

seed = ‘http://locallab-seoul.com/python/index’



tocrawl

crawled

[~~‘.../index’~~]

[]

[‘.../index’]

[‘...~~/crawling~~’,
‘.../walking’,
~~‘.../flying~~’]

[‘.../index’,
‘.../flying’]

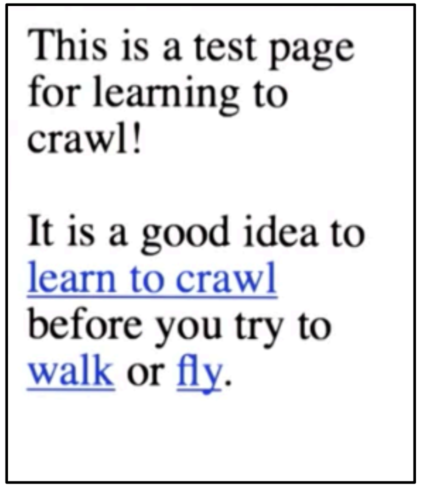
+

‘.../crawling’

+

‘.../kicking’

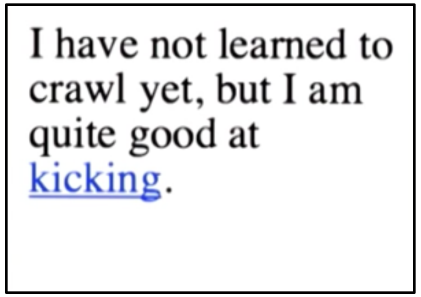
../python/index’



This is a test page for learning to crawl!

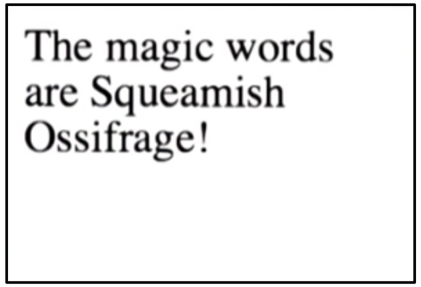
It is a good idea to learn to crawl before you try to walk or fly.

../python/crawling’



I have not learned to crawl yet, but I am quite good at kicking.

../python/flying



The magic words are Squeamish Ossifrage!

../python/kicking’

Kick! Kick! Kick!

Quiz: Crawling Process

pseudo code

```
start with tocrawl = [seed]
crawled = []
while there are more pages tocrawl:
    pick a page from tocrawl
    add that page to crawled
    add all the link targets on this page to tocrawl
return crawled
```

Quiz: Crawling Process

아래의 pseudo code 프로세스를 다음의 seed 페이지에서 시작한다면 어떤 일이 일어 나겠는가?

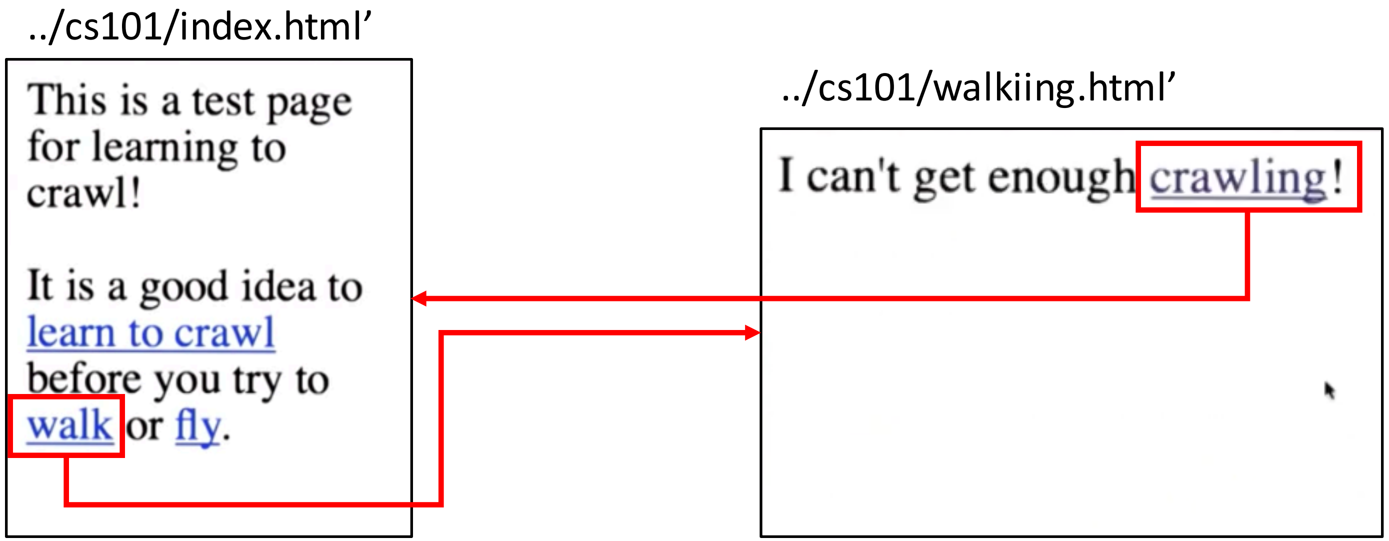
http://locallab-seoul.com/python/index

pseudo code

```
start with tocrawl = [seed]
crawled = []
while there are more pages tocrawl:
    pick a page from tocrawl
    add that page to crawled
    add all the link targets on this page to tocrawl
return crawled
```

- □ It will return a list of all the urls reachable from the seed page.
- □ It will return a list of some of the urls reachable from the seed page.
- □ It will never return.

Crawling Process



../cs101/index.html’

This is a test page for learning to crawl!

It is a good idea to learn to crawl before you try to walk or fly.

../cs101/walkiing.html’

I can't get enough crawling!

Crawling Process

```
start with tocrawl = [seed]
crawled = []
while there are more pages tocrawl:
    pick a page from tocrawl
    add that page to crawled
    add all the link targets on this page to tocrawl
return crawled
```

이미 크롤링한
페이지인지
체크

Quiz: Crawl Web

seed 페이지 주소를 입력으로하여, seed 페이지로부터 시작하여 닿을 수 있는 모든 페이지의 url을 요소로 하는 리스트를 리턴하는 crawl_web이라는 프로시저를 정의하시오.

```
def crawl_web(seed):
    tocrawl = [seed]
    crawled = []
```

Quiz: Crawl Web Loop

```
def crawl_web(seed):
    tocrawl = [seed]
    crawled =  []
    while tocrawl:
        page = tocrawl.pop()
```

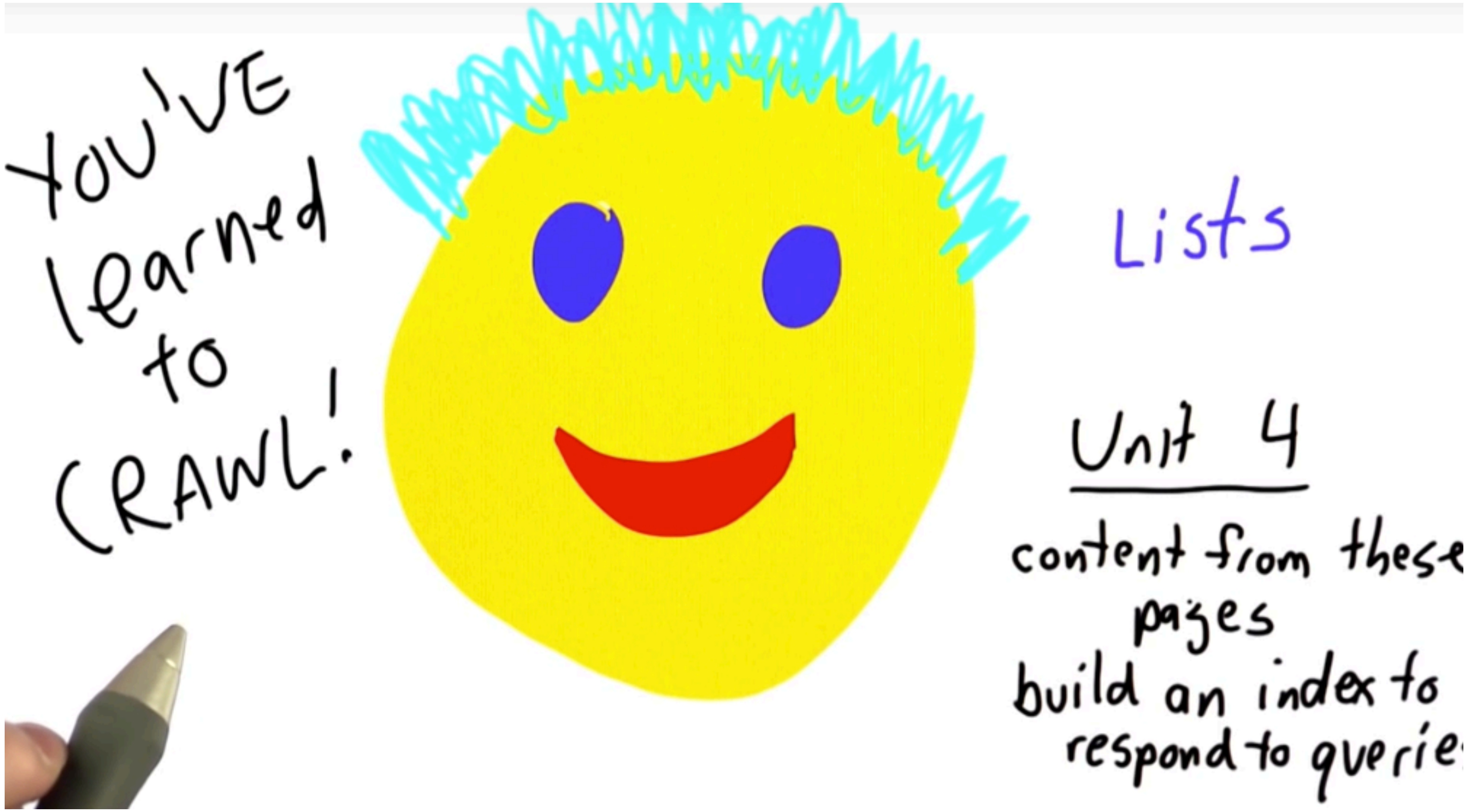
Quiz: Crawl If

```
def crawl_web(seed):
    tocrawl = [seed]
    crawled =  []
    while tocrawl:
        page = tocrawl.pop()
        if  page not in crawled  :
            crawl this page
```

Quiz: Finishing Crawl Web

```
def crawl_web(seed):
    tocrawl = [seed]
    crawled =  []
    while tocrawl:
        page = tocrawl.pop()
        if page not in crawled:
            content = get_page(page)
            union(tocrawl, get_all_links(content))
            crawled.append(page)
    return crawled
```

Conclusion



YOU'VE learned to CRAWL!

Lists

Unit 4

content from these pages

build an index to respond to queries