Introducing Web Scraping

Web Scraping

Automated extraction of data from websites; website content is first fetched (usually using HTTP) and then parsed to extract specific information.

Web Pages

Websites are collections of web pages

Web pages consist of markup e.g. HTML

This markup is understood and rendered by browsers

Fetching and Parsing

The same HTML markup can be accessed (fetched) via HTTP

Possesses an in-built hierarchical structure

Parsers can exploit this structure to extract information

Fetching Web Content

Web servers make content available on HTTP endpoints

Browsers make HTTP requests under-the-hood to get web pages

Web scraping usually involves making such requests programmatically

Many libraries and utilities available

Fetching Web Content

Command-line HTTP requests

- cURL

Python libraries for programmatic access

- Requests
- Httplib2
- Urllib, Urlllib2

Regular Expressions

Sequence of characters that define a search pattern.
Used to find strings that satisfy specific rules.

Beautiful Soup

Python package for parsing HTML and XML, including those with malformed markup such as missing tags.

Beautiful Soup

Mitigates weaknesses of regular expressions

- Global: Forms parse tree of entire HTML
- Relatively simple to use
- Robust to problems in markup being parsed

Scrapy

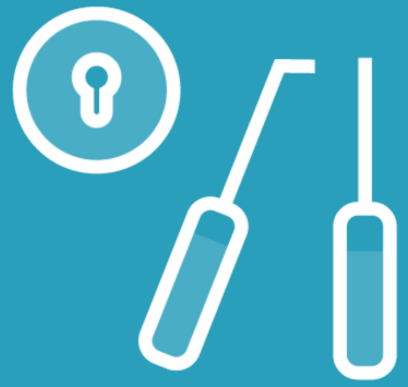
Framework for building production-grade, heavy-duty web parsing systems.

How Web Scraping Is Useful

How Web Scraping Is Useful

Profit

Web scraping business model

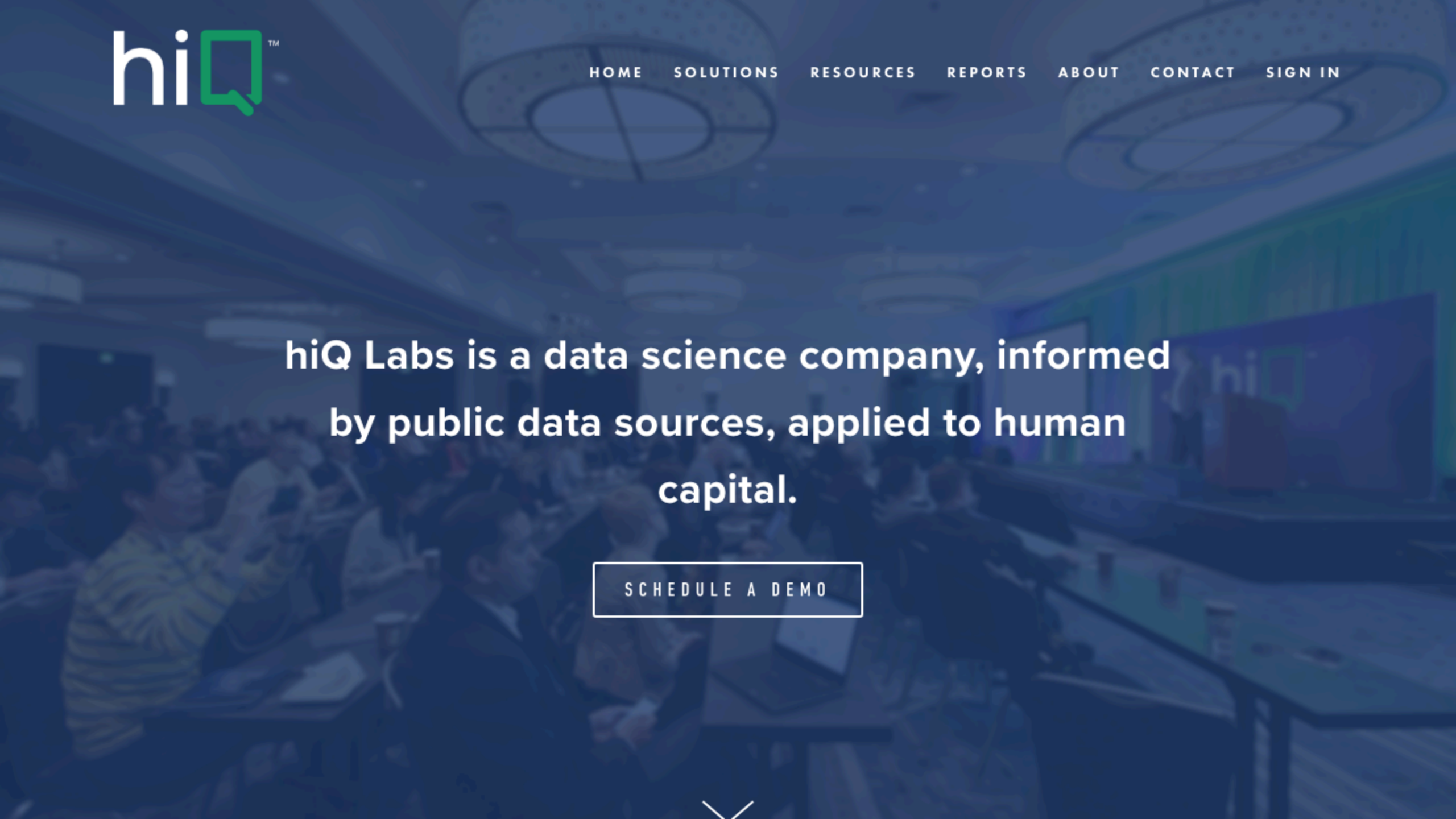


Data Access

Data is locked inside the web

Potential

Imagination + web scraping skills


hiQ™
HOME
SOLUTIONS
RESOURCES
REPORTS
ABOUT
CONTACT
SIGN IN
hiQ Labs is a data science company, informed by public data sources, applied to human capital.
SCHEDULE A DEMO


WINNER 2016
TOP HR PRODUCT
Human Resource Executive
WINNER 2016
hiQ
Dashboard for
Marketing (476)
HIGH RISK EMPLOYEES
Elizabeth Co...
Software Engineer
Discoverable: 87
Marketable: 96
Open: 85
Disengaged: 74
Pull Risk: 100
Last Action Taken: None Taken
Gary King
Internet Marketin...
Discoverable: 83
Marketable: 97
Open: 83
Disengaged: 34
Pull Risk: 100
Last Action Taken: 1 day ago
Cynthia Mora...
Internet Marketin...
Discoverable: 100
Marketable: 95
Open: 88
Disengaged: 72
Pull Risk: 99
Last Action Taken: None Taken
Michael Smith
Administrative Ex...
Discoverable: 57
Marketable: 75
Open: 78
Disengaged: 36
Pull Risk: 99
Last Action Taken: None Taken
Discoverable: 62
Marketable: 87
Open: 78
Disengaged: 46
Pull Risk:
MY ACTIONS
6
Actions Planned
4
$166M
LEAVE PROBABILITY BY EDUCATION
Probability (%)
Department Average: 10%
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
5
229
112
3
92
Associate Degree
Bachelor's Degree
Master's Degree
Doctorate
Unknown
WHY PEOPLE LEAVE
65% Of your employees are more Discoverable than the company average. These employees have professional information that is easier to find.
72% Of your employees are more Marketable than the company average. These employees are more likely to draw the attention of recruiters.
63% Of your employees are more Open than the company average. These employees are more likely to be open to external opportunities.
45% Of your employees are more Disengaged than the company average. These employees are more likely to be detached from work.
WHO WE ARE LOSING TO
LOST TO
Small (1-50)
Medium (51-500)
Large (500 - 1000)
Enterprise (1000+)
-12
ADD YOUR INTERNAL PUSH RISK INDICATORS
Refine insights and transform your organization

Hiq Labs's Competitors, Revenue, Number of Employees, Funding and Acquisitions

hiQ

Update this Profile

Get Updates

Hiq Labs's website »

hiQ designs and develops cloud-based people analytics SaaS platform for employee selection, development and retention.

Hi... Read more

Followers on Owler
140

Estimated Annual Revenue
$8.8M

Agree? Yes No

Estimated Employees
88

Agree? Yes No

CEO
Mark Weidick

CEO Approval Rating
80/100

Weigh In

OVERVIEW

Founded:	2012
Headquarters:	San Francisco, California
Status:	Private, Independent Company of
Industry Sector:	Internet Software
SIC Code:	7372 NAICS listing »
Links:	

Web Scraping Examples

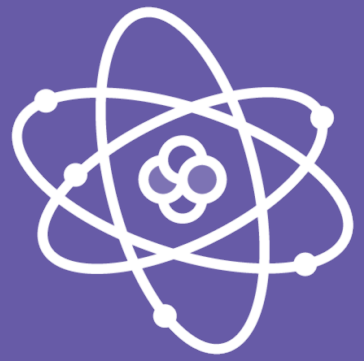
Marketing

Competitive &
Price Analysis

Lead Generation

Keyword Research

Research



Science

Product Research

Finding/Filling a Job

Government Oversight

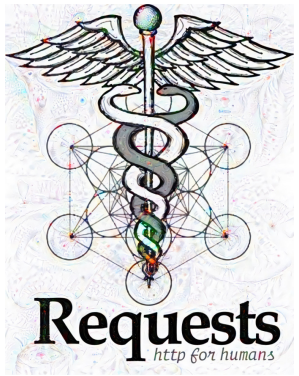
Financial

Stock Analysis

Insurance &
Risk Management

News Gathering
& Analysis

Web Scraping Python Libraries



Beautiful Soup

Requests + Beautiful Soup

Scrapy

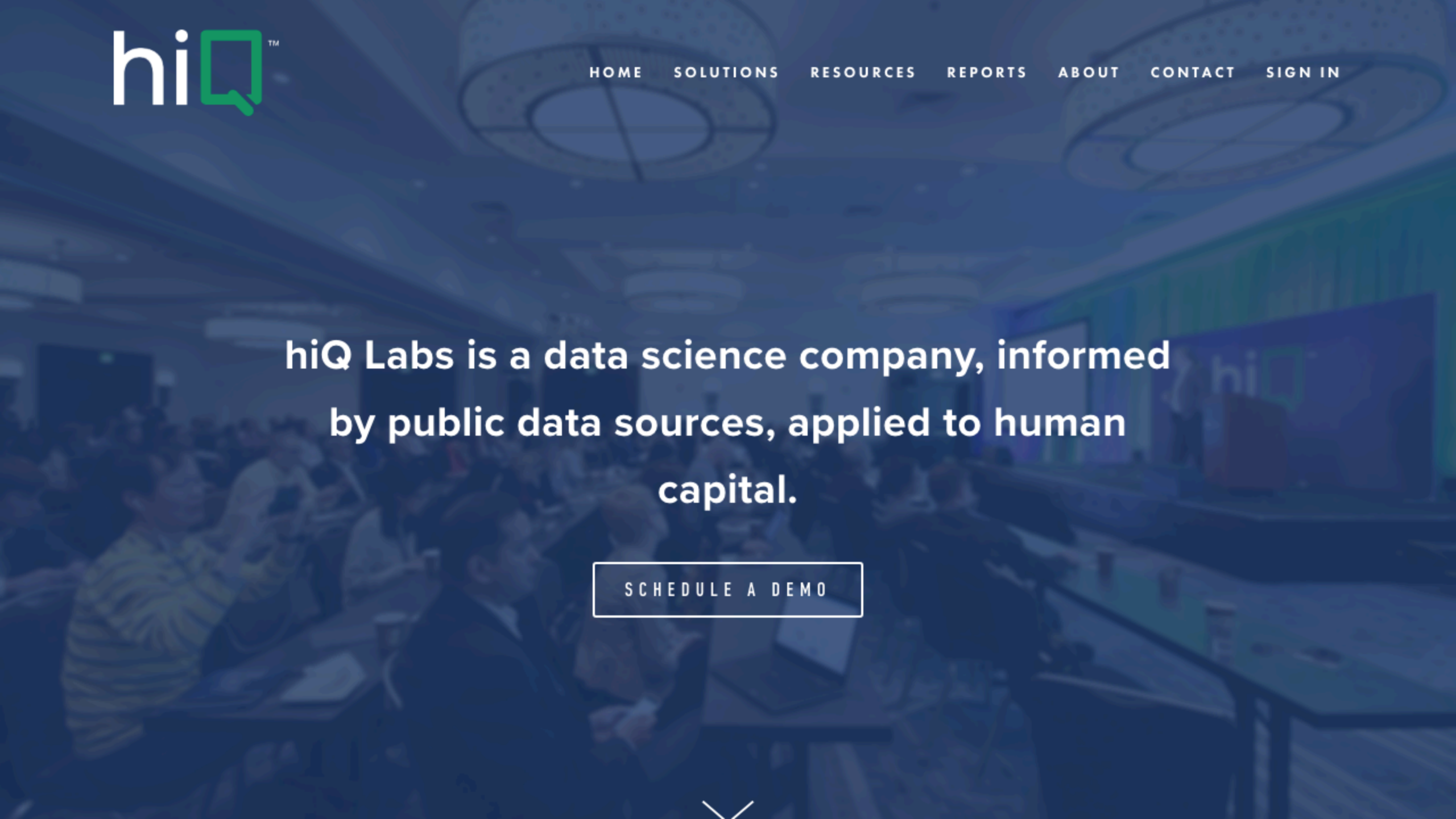
Scrapy

Selenium

Selenium

Is Web Scraping Legal? Is it Ethical?


Wiki
Blog
www
Website
Shop
advert
Content
code
optimization
Usability
link
building
Design
SEO
keywords
visits
hits
TRAFFIC
page rank


hiQ™
HOME
SOLUTIONS
RESOURCES
REPORTS
ABOUT
CONTACT
SIGN IN
hiQ Labs is a data science company, informed by public data sources, applied to human capital.
SCHEDULE A DEMO

Facts In The Legal Case

LinkedIn sent hiQ a cease-and-desist letter

Does the American anti-hacking law apply? Computer Fraud and Abuse Act (CFAA)

hiQ filed suit to protect its business

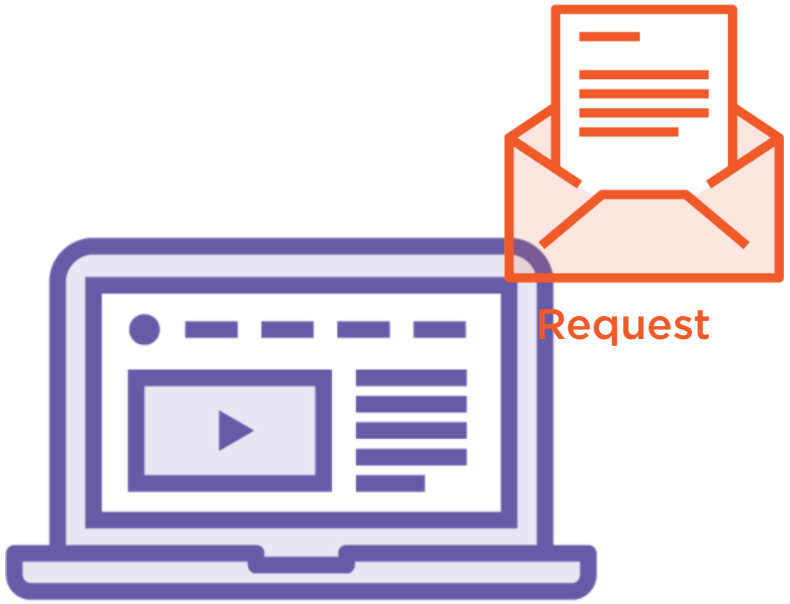
hiQ asked the court to prohibit LinkedIn from blocking access

Human Versus Web Scraper

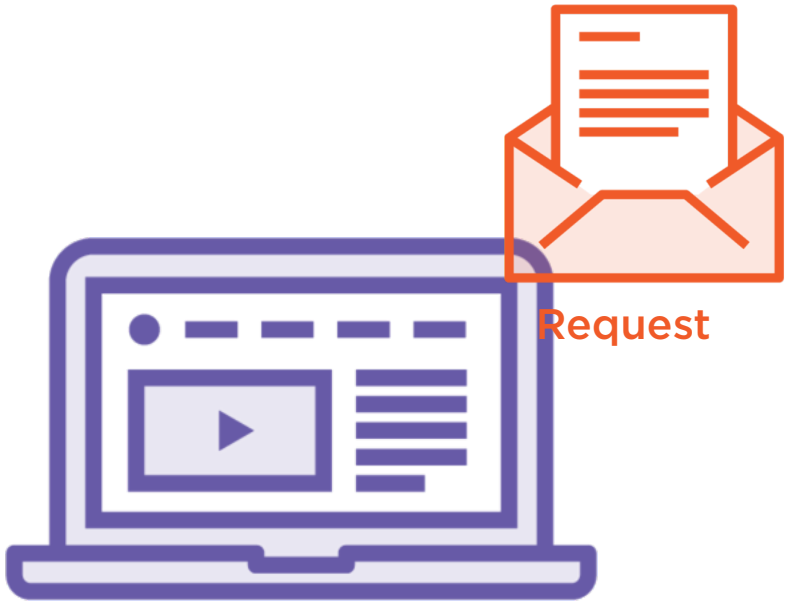
Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML
Parse HTML & render	Parse HTML
Review for Useful Information	Extract Useful Information
Interpret	Transform or Aggregate
Remember the Information	Save the Data
Click a link-Enter another URL	Go the the next URL

HTTP Overview

Request - Response



Request - Response



Hyper-Text Transfer Protocol (HTTP) is the protocol that powers the web.

HTTP Request

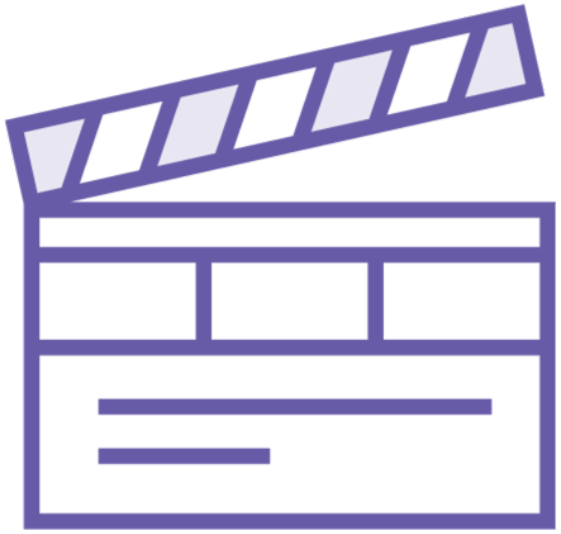
HTTP requests may include:

1 – A web address or URL

2 - A “verb”

3 - User Agent

HTTP Request: Verb



GET – Retrieves data

POST – Sends data to the server

HTTP Request: User Agent

Identifies the browser or web scraper

Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.87 Safari/537.36

URL Hacking

Used Tesla Model 3 For Sale in Lebanon, KS

Save Search

Zip Code

66952

Radius

Nation-Wide

Make

Tesla

Model

Model 3

Year

Min to Max

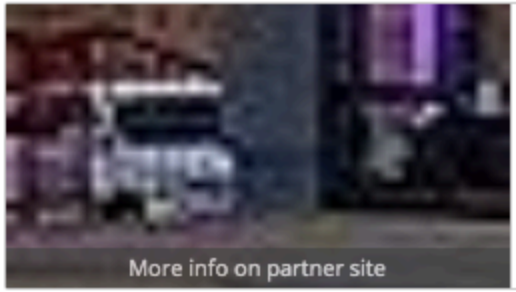
Trim

All

1-15 of 20 Used Cars Found

Best Deals First

Location:66952 X Make:Tesla X Model:Model 3 X Condition:Used X Save Search



2018 Tesla Model 3 - 22,835 mi

Ballwin, MO (435 mi) - Listed 11 days ago

$536 above market price

★★★½☆ dealer rating

1-owner , low miles , free CARFAX

$45,500

~~$45,995~~

Fair Deal

PREVIEW

SAVE

2019 Tesla Model 3 Long Range - 2,150 mi

Denver, CO (346 mi) - Listed 4 days ago

★★★★★ dealer rating

1-owner , low miles , free CARFAX

$51,995

PREVIEW

SAVE

2019 Tesla Model 3 Long Range - 2,027 mi

Denver, CO (348 mi) - Listed 2 days ago

$49,091

~~$49,690~~

PREVIEW

Used Tesla Model 3 For Sale in Lebanon, KS

Save Search

Zip Code
66952

Radius
Nation-Wide

Make
Tesla

Model
Model 3

Year
Min to Max

Trim
All

https://www.iseecars.com/used-cars/used-tesla-for-sale#Location=66952&Radius=all&Make=Tesla&Model=Model+3&Condition=used&_t=a&maxResults=15&sort=BestDeal&sortOrder=desc&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D

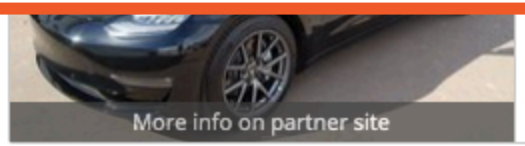
1-15 of 20 Used Cars Found

Best Deals First

Save Search

$45,500
~~$45,995~~
Fair Deal
PREVIEW
SAVE

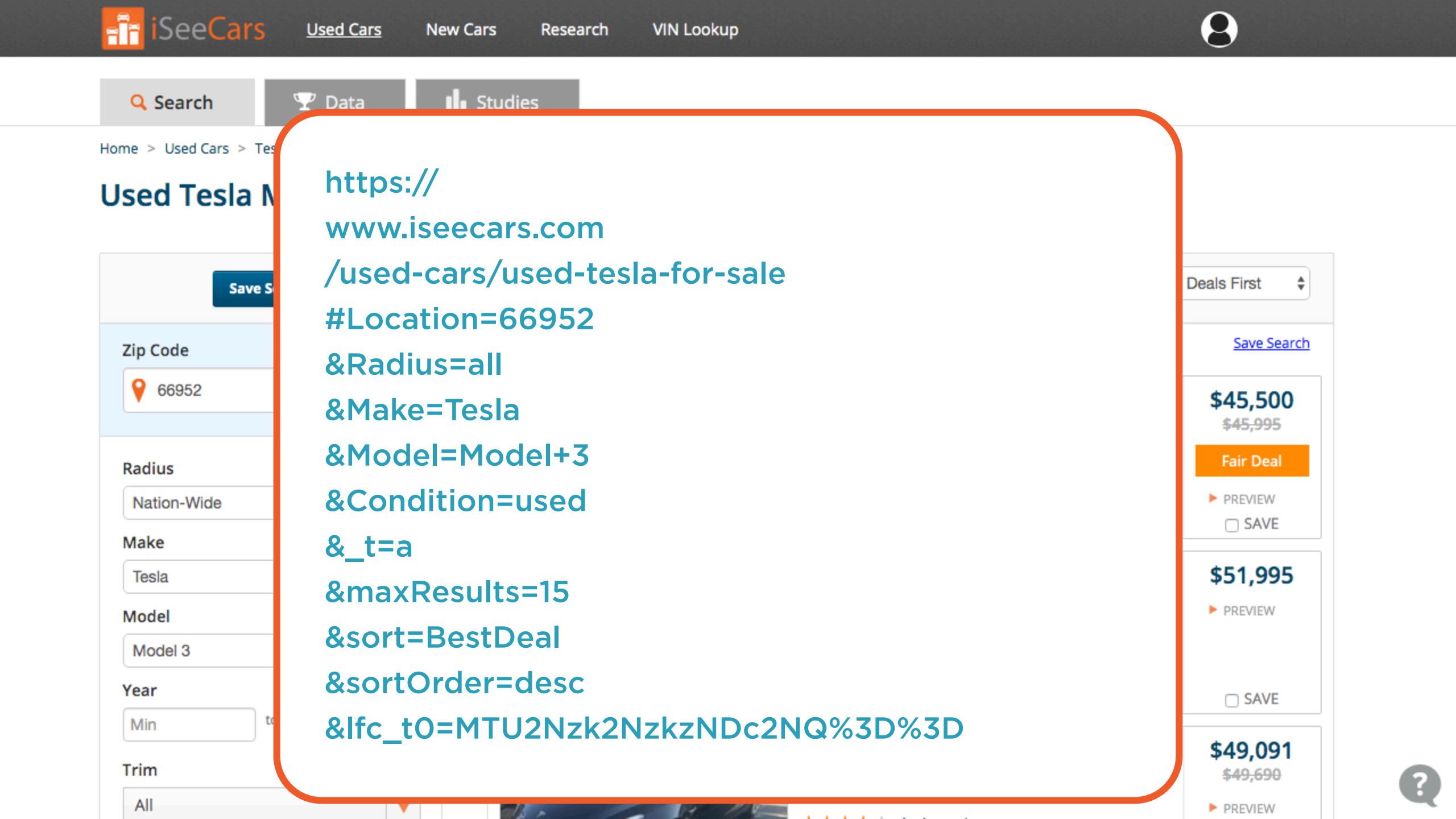
$51,995
PREVIEW
SAVE



★★★★★ dealer rating
1-owner , low miles , free CARFAX

2019 Tesla Model 3 Long Range - 2,027 mi
Denver, CO (348 mi) - Listed 2 days ago

$49,091
~~$49,690~~
PREVIEW


iSeeCars
Used Cars
New Cars
Research
VIN Lookup
Search
Data
Studies
Home > Used Cars > Tes
Used Tesla M
Save S
Zip Code
66952
Radius
Nation-Wide
Make
Tesla
Model
Model 3
Year
Min
Trim
All
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
Deals First
Save Search
$45,500
$45,995
Fair Deal
PREVIEW
SAVE
$51,995
PREVIEW
SAVE
$49,091
$49,690
PREVIEW

Scheme

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Host

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Port

iseescars.com URL

```
https://
www.iseecars.com:443
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Path

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String (?)
or
URL Fragment (#)

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

```
host = ‘www.iseecars.com'
path = ‘/used-cars/used-tesla-for-sale’
location = ‘66952’
query_string = f’#Location={location}&Radius=all&Make=Tesla&Model=Model+3’

start_url = f’http://{host}{path}{query_string}’
```

Python URL Strings

```
import requests
start_url = ‘https://www.iseecars.com/used-cars/used-tesla-for-sale’

downloaded_page = requests.get(start_url)

print(downloaded_page.text)
```

Python Requests