Exploring Web Scraping with Python

WHY SCRAPE THE WEB?



Clarke Bishop
BIG DATA ENGINEER

@ClarkeBishop www.clarkebishop.com

Overview

How web scraping is useful

Web scraping examples

Scraping for a car

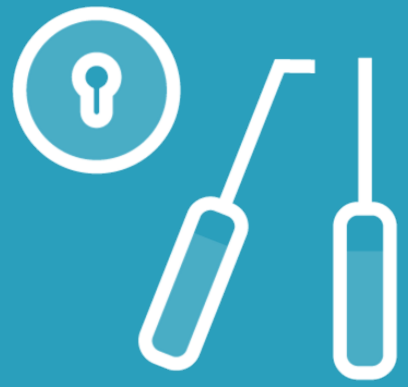
How Web Scraping Is Useful

How Web Scraping Is Useful



Profit

Web scraping business model



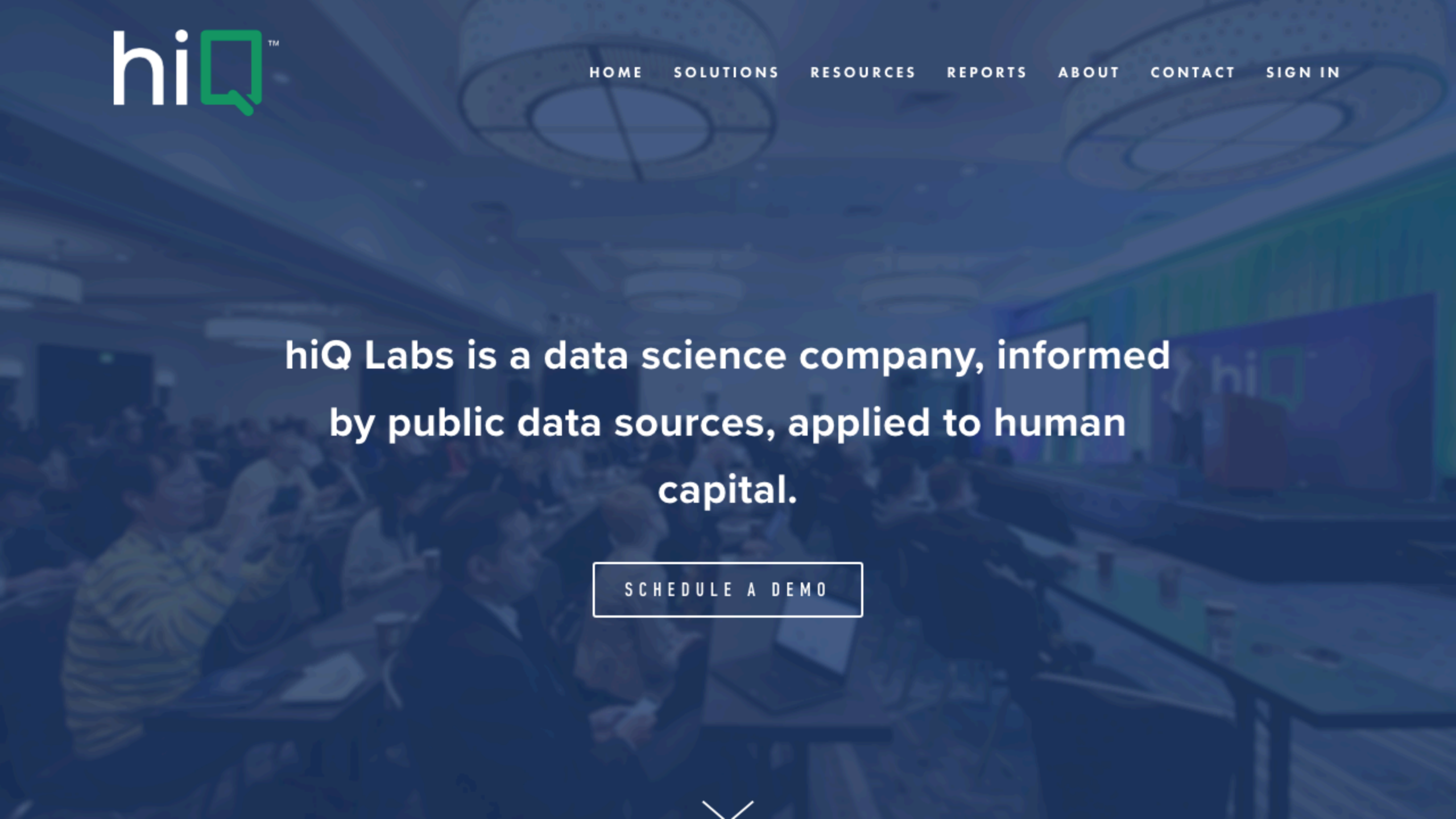
Data Access

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



Clarke Bishop

CTO ✦ Cloud Solutions Architect ✦ Engineer ✦ Problem Solver

Roswell, Georgia · 500+ connections

Join to Connect

Client Command

Georgia State University - J. Mack Robinson College of...

Websites

About

Fixing the Innovation Dilemma = Operational Leadership to Solve the Adoption of Innovation Problem.

▶ WHAT'S BROKEN: Innovation is hard. Driving the adoption of innovation—both internal alignment and to the marketplace is harder.

People also viewed



Uday kumar P
UI Developer at Delta Air Lines



Emily Cade
Assistant Manager at Buckle



Gregory Lawson
Cloud Solutions Architect at Global Payments Inc.

Emily Emerson



Albert (Bert) Sams
Vice President of Information Technology, Haggai International

Madhavi Patel
Project Manager/Business System Analyst at Santander Bank, N.A.



Ryan Gunnigle
Kids2 CEO (baby toy & gear), Be Curious Partners Co-Founder (FamTech VC), Hello Einstein Studios Co-Creator (Media)

Hiq Labs

Overview
Competitors
News
Press Releases
Videos
Contact
Summary
FAQ

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

hiQ

Update this Profile

Get Updates

Followers on Owler
140

Hiq Labs's website »
hiQ designs and develops cloud-based people analytics SaaS platform for employee selection, development and retention.

Hi... Read more

Estimated Annual Revenue
$8.8M

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:	

Estimated Employees
88

Agree? Yes No

Web Scraping Examples

Marketing



Competitive &
Price Analysis

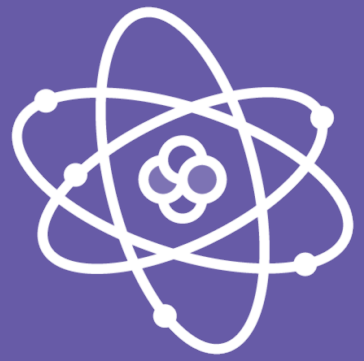


Lead Generation



Keyword Research

Research



Science



Product Research



Finding/Filling a Job



Government Oversight

Financial



Stock Analysis



Insurance &
Risk Management



News Gathering
& Analysis

Unleash Your Imagination

Scraping for a Car



Scrape Up a Tesla

Used Teslas for Sale in Atlanta, GA

Save Search

Used Cars | New Cars

Location: Atlanta, GA
Distance: 25 miles
Make: Tesla
Model: All
Body Style: All
Years: Min to Max

Price
$0 – $100,000+

Mileage
0 – 200,000+

Showing 1 – 30 of 62 Listings

Sort By: Best Match

Featured Results



2016 Tesla Model S 70D AWD

Mileage: 48,036 miles
Distance: 24 mi. - Lawrenceville, GA
Exterior: Black
Interior: Black

Good Price
$45,410

All Results



2013 Tesla Model S Base RWD

Mileage: 71,085 miles
Distance: 6 mi. - Atlanta, GA
Exterior: Black
Interior: Black

Exceptional Price
$31,200

Tesla Buying Adventure







Scraping Up a Tesla

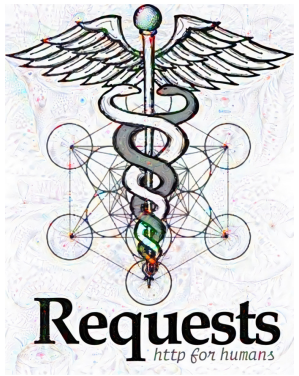
Scrape the car buying websites to find all the Teslas

Evaluate the prices to find truly great deals

Scrape airfares and adjust the deals for airfare expenses

Send an email digest of the top deals each day

Web Scraping Python Libraries



Beautiful Soup

Requests + Beautiful Soup

Scrapy

Scrapy

Selenium

Selenium

Summary

Why scrape the web?

The web is filled with profitable data

For data trapped in web pages, scraping is the only option

Scraping can make it easier to buy a car