



# The Proverbial Fork in the Road

Your organization, like many, needs information from the Internet in order to make timely and accurate business decisions. Your job is choosing the best way to get that information and make it usable for a wide range of business processes. Basically, you have two ways you can go – web scraping or web information integration.

This guide will show you what the future for you and your organization could be like when you opt for web scraping versus web information integration.

 What you do after you've reviewed it is your decision.  
*(Just don't say we didn't warn you.)*



*Pick the right path  
and you can anticipate  
a brighter future for you  
and your company.*

*Pick the wrong path  
and that future could  
be overcast by uncertainty  
and doubt.*





# Web Scraping Software

At first glance, using web scraping tools seems like a no-brainer. Inexpensive to purchase or able to be scripted in-house using existing resources, start-up costs are minimal. Web scraping is a simple way to extract data from a website where no programming interface exists. You just tell it what you want and it goes and gets it. *What could possibly go wrong?*



## What Could Possibly Go Wrong

-  Every step of the way, you're dogged by development problems, programming errors and requests for things web scraping can't do.
-  You don't get the data you need (or not all of it, anyway) and the data you do get, you don't need (at least not over and over again).
-  Development costs start piling up and the need to clean up data delays putting it to use, resulting in lost revenue opportunities.
-  Websites change (imagine that!) and you discover web scraping software has some serious issues when it comes to dealing with change.



## On Second Thought

*If web scraping doesn't meet everyone's expectations – or worse yet – creates more problems than it solves, what initially seemed like a cakewalk could turn into a minefield.*



# Web Information Integration

does a lot more than web scraping. It digs deeper into the web and eliminates time-consuming and costly manual processes needed to verify information and ensure its relevance. It also automates the entire process of data acquisition and integration with business systems and processes to enable almost real-time, data-driven decision making. *What could be better?*

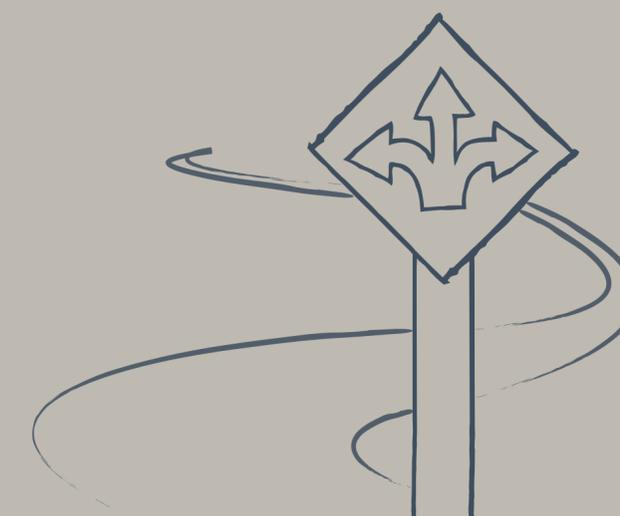


## What Could Be Better?

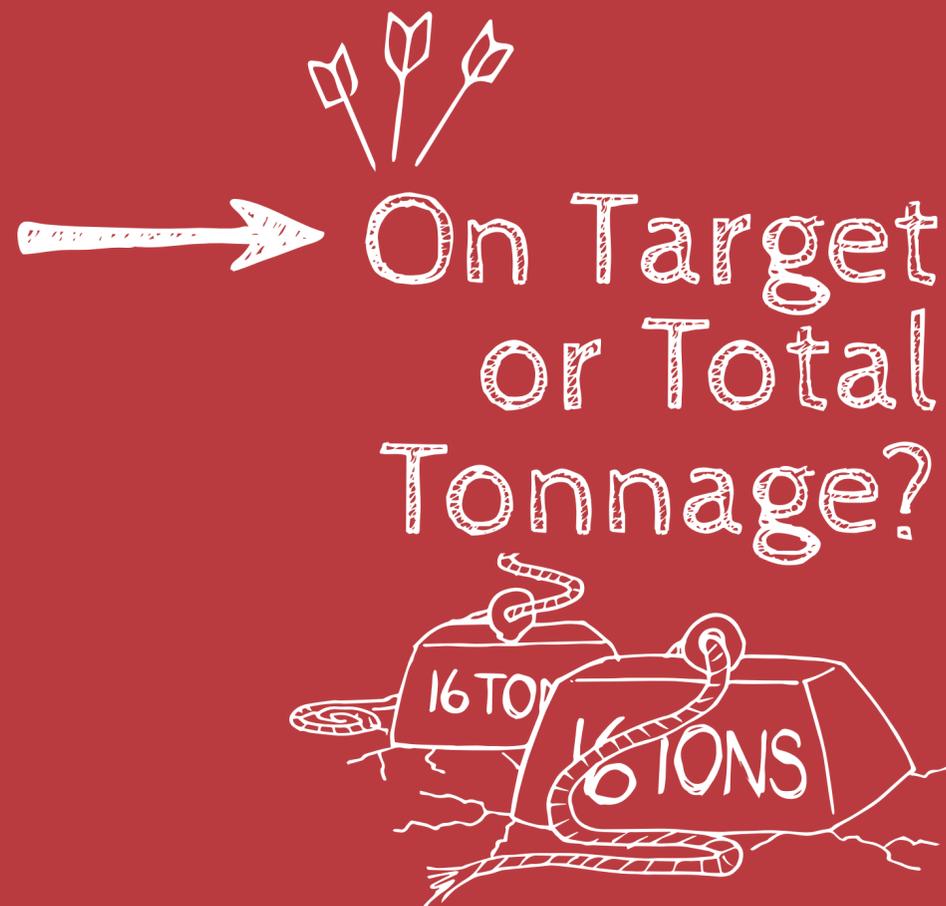
- 👍 Acquiring comprehensive, actionable business information from the web automatically and continuously.
- 👍 Seamless, effortless and nearly instantaneous integration of web information with business processes.
- 👍 A return on an initial investment that increases over time by saving your company money and making it more competitive.
- 👍 Your recognition as “*Employee of the Month*” followed by perks, promotions – and most importantly – power.

Better   
Think Again

Since web information integration is truly a game changer, maybe you should reconsider whether web scraping is the best path for you and your company to set out on.







Before you dive in the deep end of the Internet data pool, you have to ask yourself,



*“What the heck am I doing this for?”*

Because if you just want to gather lots of information from the Internet – regardless of whether it’s inaccurate or redundant – then **web scraping** can probably get the job done.

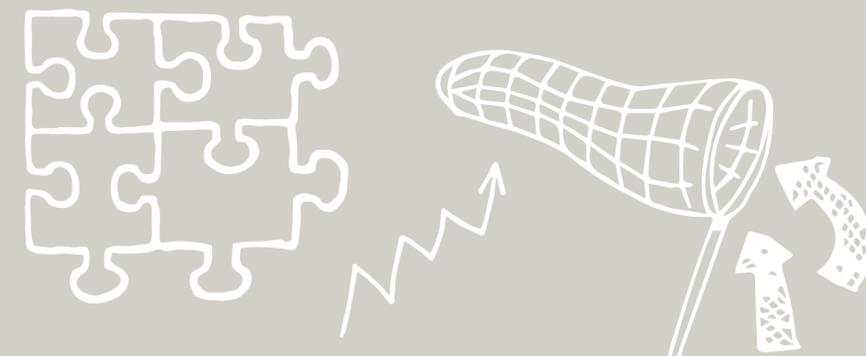
But if you want to extract precisely the data you need – and feel confident it can be used for making business decisions – then **web information integration** is a better way to go.



Complete or  
Incomplete?

Web scraping acquires some of the data, but never gets it all.

It doesn't support data acquisition on dynamic scripts like AJAX and JavaScript and hits a wall with dynamic websites and complex extraction routines. Reformatting or even minor changes to a website can also stop it dead in its tracks.



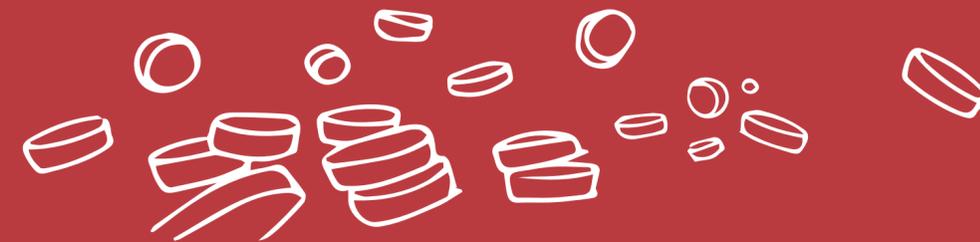
Web information integration is – in a word – complete. Comprehensive. Thorough. All-inclusive. Exhaustive. The whole nine yards.

(OK, that's more than one word, but you get the picture.) It casts a wider net and scours the web for data from sites that include partner portals and public online data verification services as well as public domains.

100%

Immaculate or Inaccurate?

WRONG

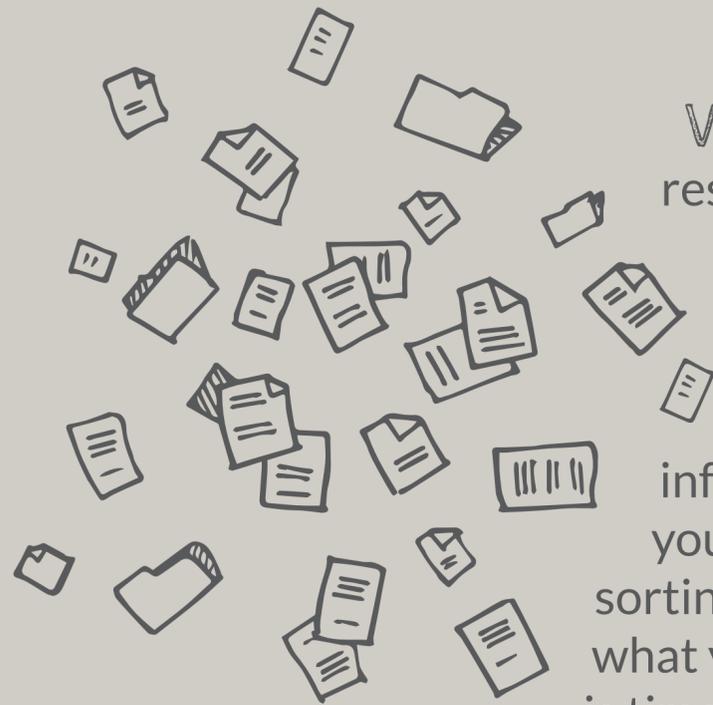


Information acquired by **web scraping** software can be inaccurate or unusable. Data might not be in the right format, for example, or it might not be refreshed, rendering it out-of-date.

**Web information integration** extracts data from the Internet, scrubs it squeaky clean and enriches it by eliminating extraneous noise and combining it with data from different sources.



# Irreverent or Relevant?



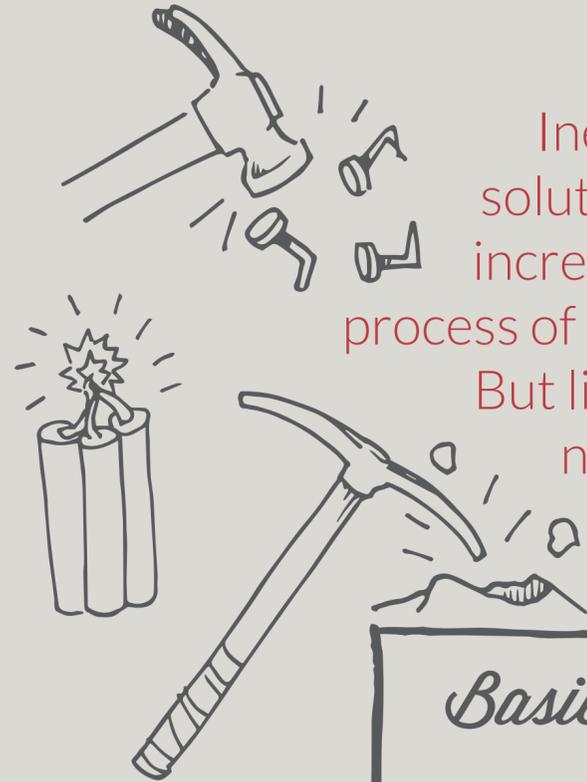
Web scraping doesn't respect boundaries when it comes to collecting data and it can't stay on top of website changes to filter out unwanted information either. So what you see is what you get, and sorting through it all to deliver what you really want and need is time-consuming and tedious.

Web information integration eliminates unwanted data by extracting only what you need from the Internet. It then integrates that data with a host of enterprise systems, databases, data warehouses and business intelligence tools, turning it into information that's relevant and easy to use.





# The Awe-Encompassing Question



Inexpensive **web scraping** solutions are crude tools for the increasingly complex and critical process of gathering web information. But like physical scrapers, they're neither 100 percent effective nor 100 percent accurate.

**Web information integration** is a precision-guided approach to extracting relevant data from the Internet and transforming it into actionable information that can immediately increase the competitiveness of your company.



*Basically it comes down to a question of quantity vs. quality.*

(Hint: "Quantity" is the wrong answer.)



-----> It's a  
Matter of  
Your Time



You could give **web scraping tools** a try and by the time their expense and inadequacy catch up with you, you could find another job.



or



You could implement **web information integration** now and start taking credit for business intelligence that sets your company apart from the competition right away.





## Getting Started



With web scraping, one size doesn't fit all.

An entry level tool might be modified to work initially, but ongoing front-end development would require copious amounts of time-consuming coding and a plethora of programmers to keep it up and running. You could also custom build your own solution, but that would cost more and take even longer before you could start dredging the web.



Web information integration can be deployed within days or even hours vs. weeks or months with web scraping tools.

Featuring a non-coding interface for non-technical users, it automates the manual operations a human would do when acquiring data from the web. Simply work your way through a site the way you normally would, and the data extraction process is automatically built. Hit "run" and it navigates a site and extracts the data just as you told it to do.



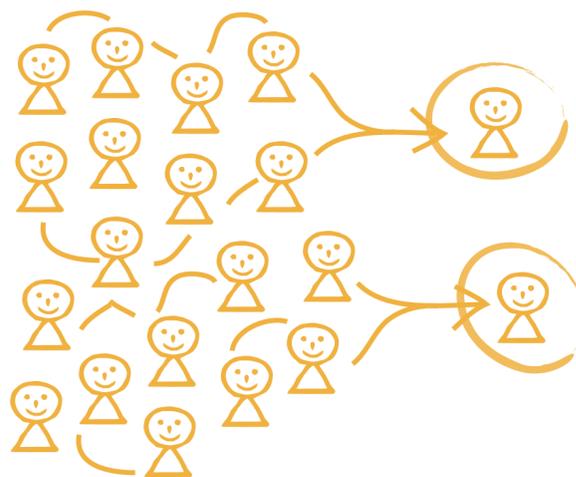
## ➔ Down Time



### Web scraping tools

aren't designed for mission-critical business applications – so they're not very resilient and constantly break down. They can require reprogramming every time websites change or get reformatted – which happens, on average, every 18 months. Of course, you can only tell they've been reformatted by detecting gaps in data – and who knows how long it will take for someone to spot those?

---



While developers can still write code and interface directly with systems if they need to, **web information integration** is a robust solution designed to serve the enterprise. With actionable information that avoids the added expense of maintenance and reprogramming. By switching from “inexpensive” web scraping to web information integration, for example, one company decreased its number of developers from 35 to two.

# → Lag Time



With **web scraping**, it takes a long time to get from raw data collection to real data comprehension. First, someone has to sort through all the information that's acquired to verify what's accurate and relevant. Then they still have to process it before it can be used for making business decisions.



**Web information integration** makes actionable information available almost in real-time. It collects and cleans the data and automatically integrates it with a host of enterprise systems and business intelligence tools where it can quickly be accessed and immediately acted on by users.





## Time to Switch

You may have assumed setting up **web scraping** was something your IT research analyst team could handle. Or that the added cost of overtime and hiring more programmers would fly under the radar. And the hidden costs of delays, downtime and decreased competitiveness might go unnoticed. So, how lucky do you feel?



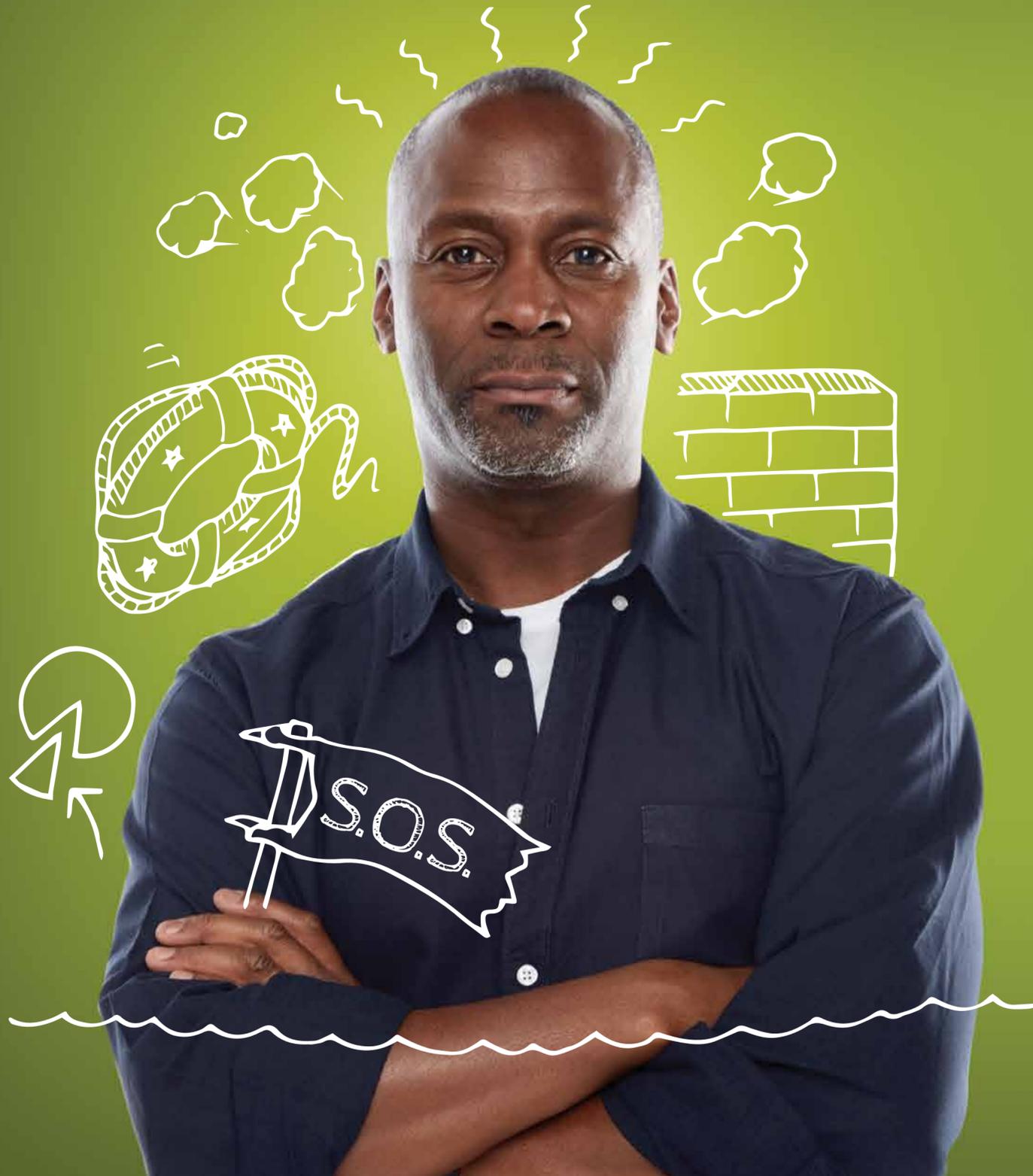
Fortunately, it's not too late to back out

After all, you've heard of one-click programming before. What if it's true this time? It's also possible delivering data on time to the front lines might impress the suits and get you invited to a power lunch or two. Or that keeping the lid on overtime and new hires could make you a hero with the front office and give you a step up the corporate ladder?

That alone should cause you to pivot on the path you were headed down, but just to make sure,  
*check out one more reason.*

Future A:

Perpetual frustration, abject humiliation and the answer to “What else could go wrong?”



Future B:

“Sky’s the limit” scalability, unbridled expectations and your key to the executive wash room.



## ⇒ Ramp Up or Shut Down



Let's say your **web scraper** tool is sputtering along with nominal results and someone decides that's good enough to double its capacity. You can't go there.



*Seriously, you can't go there.*



However, if you've been paying attention and opt for **web information integration**, there's no limit to the heights you and your organization can aspire to. Just go for it.

## Size Matters



Web scraping is a static approach to a dynamic environment.

The more sites you add or the more you increase the volume or frequency of acquisitions, the more likely it is to come crashing down.

Web information integration is infinitely scalable.

It can collect and combine data from a multitude of disparate sources – both internal and external – and add new sources without skipping a beat.





# Flexible Is as Flexible Does



Designed primarily for individual users, **web scraping** can only perform one-way data acquisition. It has few — if any — built-in features capable of integrating information based on business needs.



Data acquisition is only a small part of what **web information integration** can do. It can also take that data and integrate it with business applications and processes accessible throughout an organization. And do it all automatically and almost instantaneously.



# → When Data Means Business

While error-prone and incomplete, web scraping does eliminate some manual processes required for collecting data from multiple websites,

that's pretty much it.



Web information integration can leverage data to satisfy the requirements for almost any data-driven intelligence application that any organization – from the smallest company to the largest multi-national corporation – wants to do. So whether it's competitive price monitoring, sales intelligence, news aggregation, background checks, brand surveillance, financial investment research, IP protection and fraud detection, OSINT and more, it's got you covered.



## → What Were You Thinking?

Maybe you thought that you were good to grow with **web scraping**. Or that by the time you were asked to scale it, you'd have moved on.

So much for thinking.

Now you can see the difference as clear as day. Web information integration is scalable. Web scraping isn't.



That's because web scraping was intended to do one thing and one thing only – acquire boatloads of unverified information.

Web information integration, however, was designed with a loftier goal in mind – giving users unlimited access to relevant data-driven processes that drive a company's business and give it a competitive edge both now and in the future.

Future A:

Data deficient, hopelessly disconnected and willing to consider a career change.



Future B:

Validated, integrated, automated and brimming with swagger and self confidence.



# ➔ There Will Be a Test

So far, the differences between web scraping and web information integration may have seemed purely academic. But things start getting real, real fast, once either of them gets deployed. So let's take a look at what life might be like after that.

## Web Scraping

- ✓ Implementation encounters endless delays and interminable round table discussions. The word around the water cooler is it's not going to happen.
- ✓ Once it's deployed, coworkers are required to spend endless hours manually processing data to ensure its validity and relevance. They're not happy.
- ✓ Much of the data is incomplete and inaccurate making it — in a word — unusable. As a result, interest in web scraping declines precipitously.
- ✓ As the realization sinks in that web scraping is a dead-end solution, the organization starts considering alternatives. You're not part of those discussions.
- ✓ There's a growing concern throughout the company it could lose its competitive edge without actionable web intelligence. Not so much about losing you.

## Web Information Integration

- ✓ All human interactions previously required to extract and organize data from websites have been automated – and on a scale previously unimaginable.
- ✓ Relevant data derived from every targeted website is seamlessly merged with internal systems and processes and is almost immediately available for use in making business decisions.
- ✓ Adding sites or increasing the volume of data is effortless, making it easy to keep pace with the growing demand for web information integration services.
- ✓ Awareness of web information integration's value keeps increasing as its cost of ownership declines and the ROI keeps going up and up.
- ✓ Coworkers admire you and the company brass holds you in the highest esteem. Business is good. Sales are good. You're good. It's all good.

To learn more about how web information integration can  
create a more promising future for you and businesses  
like yours, contact us today at  
[kofax.com/web-data-extraction](http://kofax.com/web-data-extraction)

**KOFAX** 

Contact in Poland:  
**Robotic Competence Center**  
**Rhenus Data Office Polska Sp. z o.o.**  
A: al. Katowicka 66, 05-830 Nadarzyn  
T: +48 (22) 609 280 280  
E: [bartosz.kochanowski@pl.rhenus.com](mailto:bartosz.kochanowski@pl.rhenus.com)  
W: <http://www.robotydooroboty.pl>  
W: <http://www.hellorobotics.pl>

