


# Introducing Scrapy

---

## Scrapy

Scrapy is an application framework for crawling web sites and extracting structured data

Originally built for web scraping  
but now used for web crawling



## Scrapy

Scrapy is an application framework for crawling web sites and extracting structured data

# Scraping vs. Crawling

## Web Scraping

Extract data directly from web sites

Data analysis and somewhat  
unsavory reputation

Specific – “scrape prices from Amazon”

Small scale, results in specialized dataset

## Web Crawling

Download and index web sites

Performed all by search engines,  
associated with legitimate use

General – “crawl sites linked off Amazon”

Large scale, results in document corpus

Framework vs. library:  
inversion of control

## Scrapy

Scrapy is an **application framework** for crawling web sites and extracting structured data



# Library vs. Framework

## Library

You call library functions

You write the application and invoke library for specific portions

## Framework

Framework calls you

Framework defines the application and invokes your code for specific portions

Beautiful Soup is a parsing library

Scrapy is a web scraping framework

You must know what you are  
looking for – tied to **HTML** format

## Scrapy

Scrapy is an application framework for crawling web sites and extracting  
**structured** data





Inherently somewhat **fragile**, like  
regular expressions and other  
related tools

## Scrapy

Scrapy is an application framework for crawling web sites and extracting  
**structured** data



Specific HTML elements are selected  
for processing using **Selectors**

## Scrapy

Scrapy is an application framework for crawling web sites and extracting  
**structured** data



Scrapy supports selectors  
specified in CSS and XPath

## Scrapy

Scrapy is an application framework for crawling web sites and extracting  
structured data

data

## **Selector**

Specification of what HTML elements ought to be selected for processing. Scrapy supports XPath and CSS selectors.