Selecting Elements Using the Scrapy Shell

Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

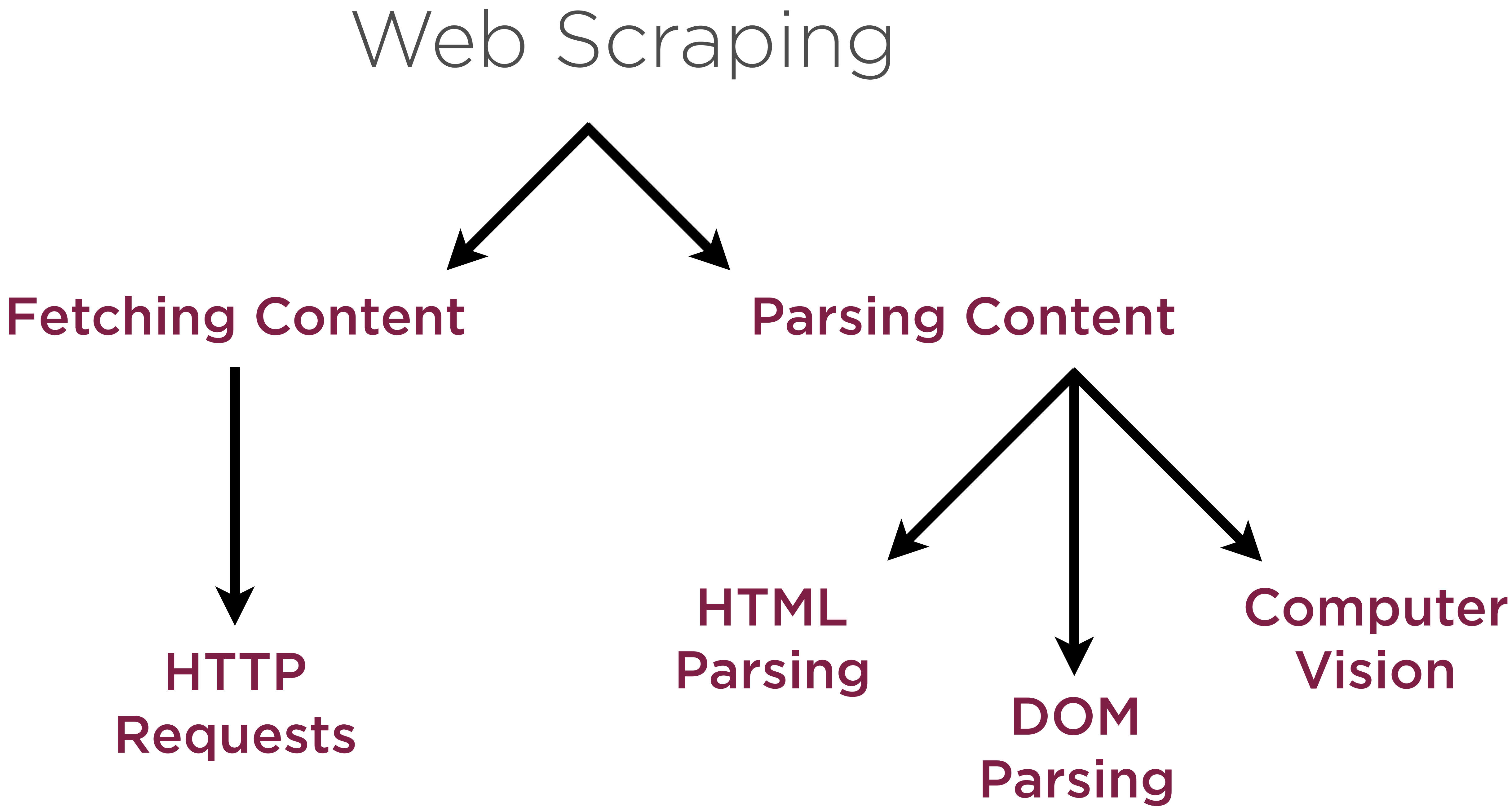
Scrapy as an application framework for crawling websites

Data extraction in a structured format

Interactively testing extraction using the Scrapy shell

Leveraging XPath and CSS selectors in the Scrapy shell

Parsing Web Content



Web Scraping

Fetching Content

Parsing Content

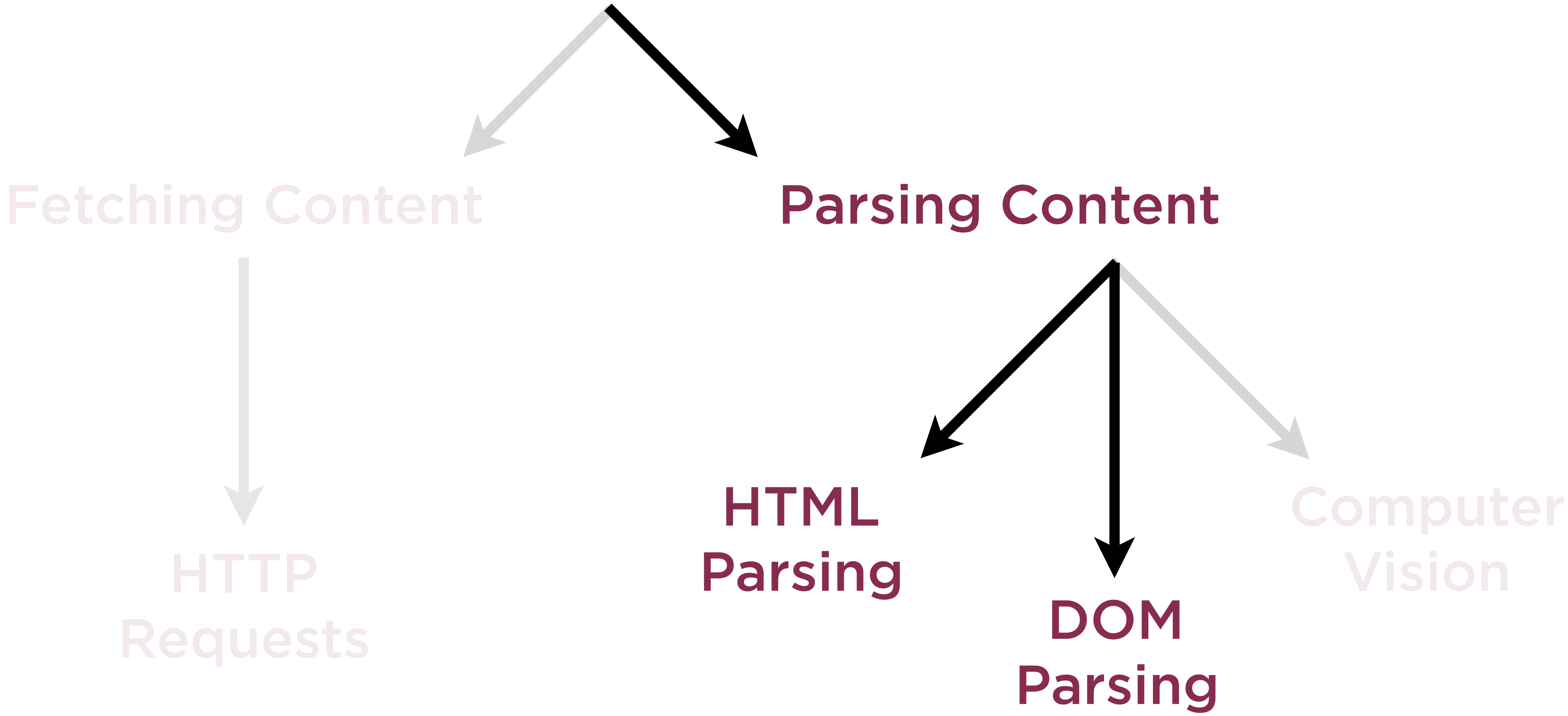
HTTP Requests

HTML Parsing

DOM Parsing

Computer Vision

Web Scraping



Fetching Content

Parsing Content

HTTP Requests

HTML Parsing

DOM Parsing

Computer Vision

Parsing Web Content

HTML Parsing

Parse HTML and CSS associated with web content

Can not parse dynamic changes made by Javascript

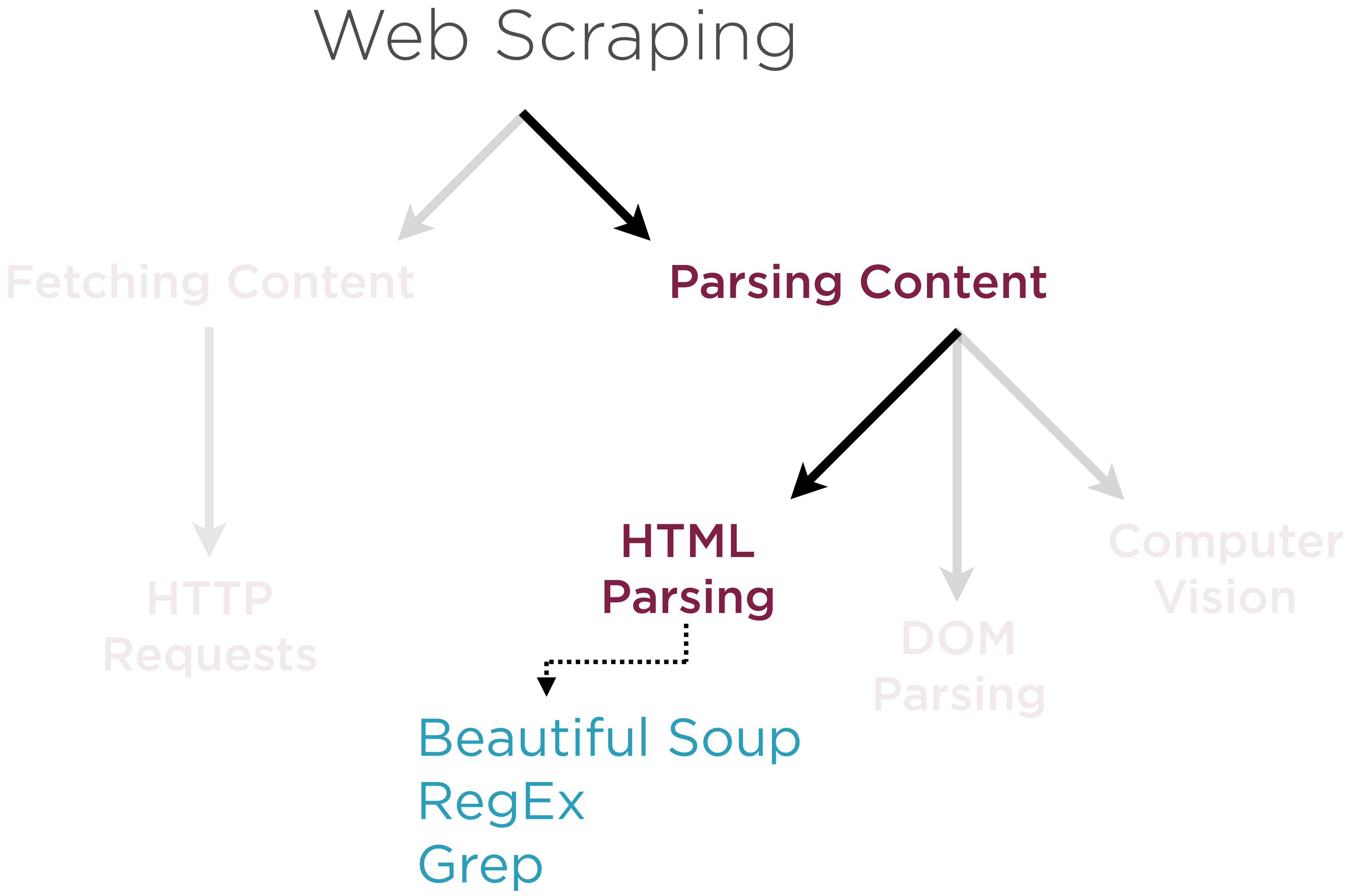
Work with structured text, rather than object of the Document-Object-Model

DOM Parsing

Parse dynamic content in addition to static content

Can parse dynamic changes made by code elements

Can work either with text elements or with objects in the DOM



Web Scraping

Fetching Content

Parsing Content

HTTP Requests

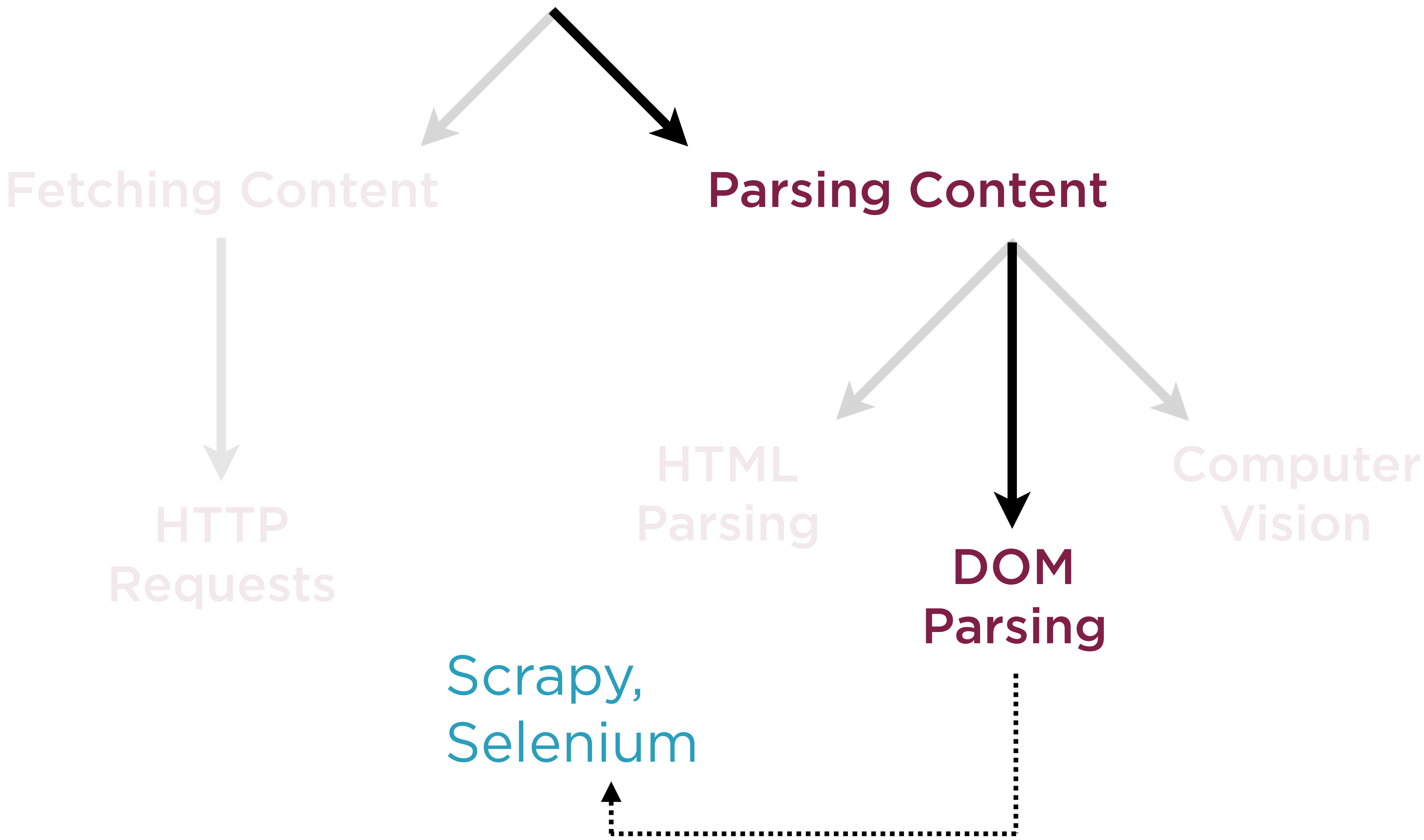
HTML Parsing

DOM Parsing

Computer Vision

Beautiful Soup
RegEx
Grep

Web Scraping



Fetching Content

HTTP Requests

Parsing Content

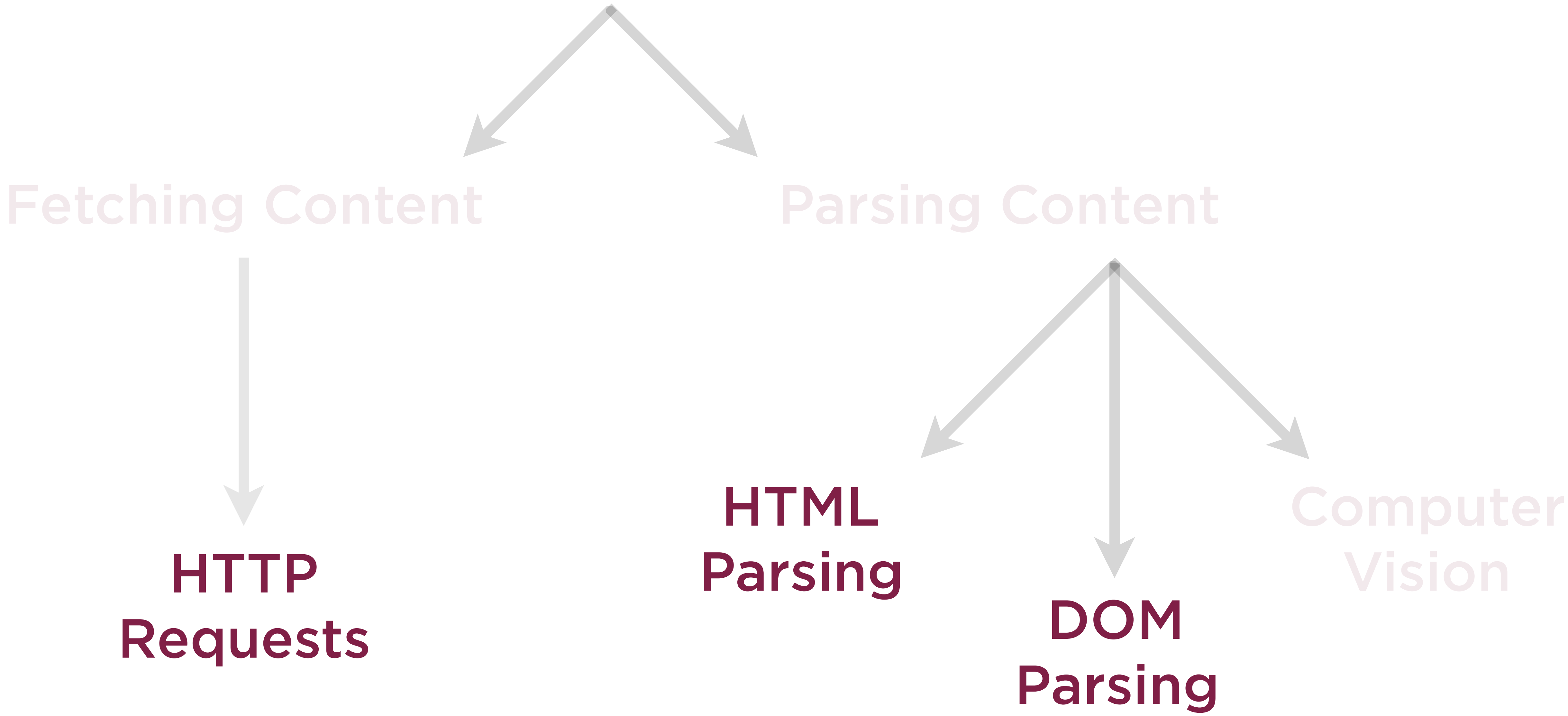
HTML Parsing

DOM Parsing

Computer Vision

Scrapy, Selenium

Web Scraping



Fetching Content

Parsing Content

HTTP Requests

HTML Parsing

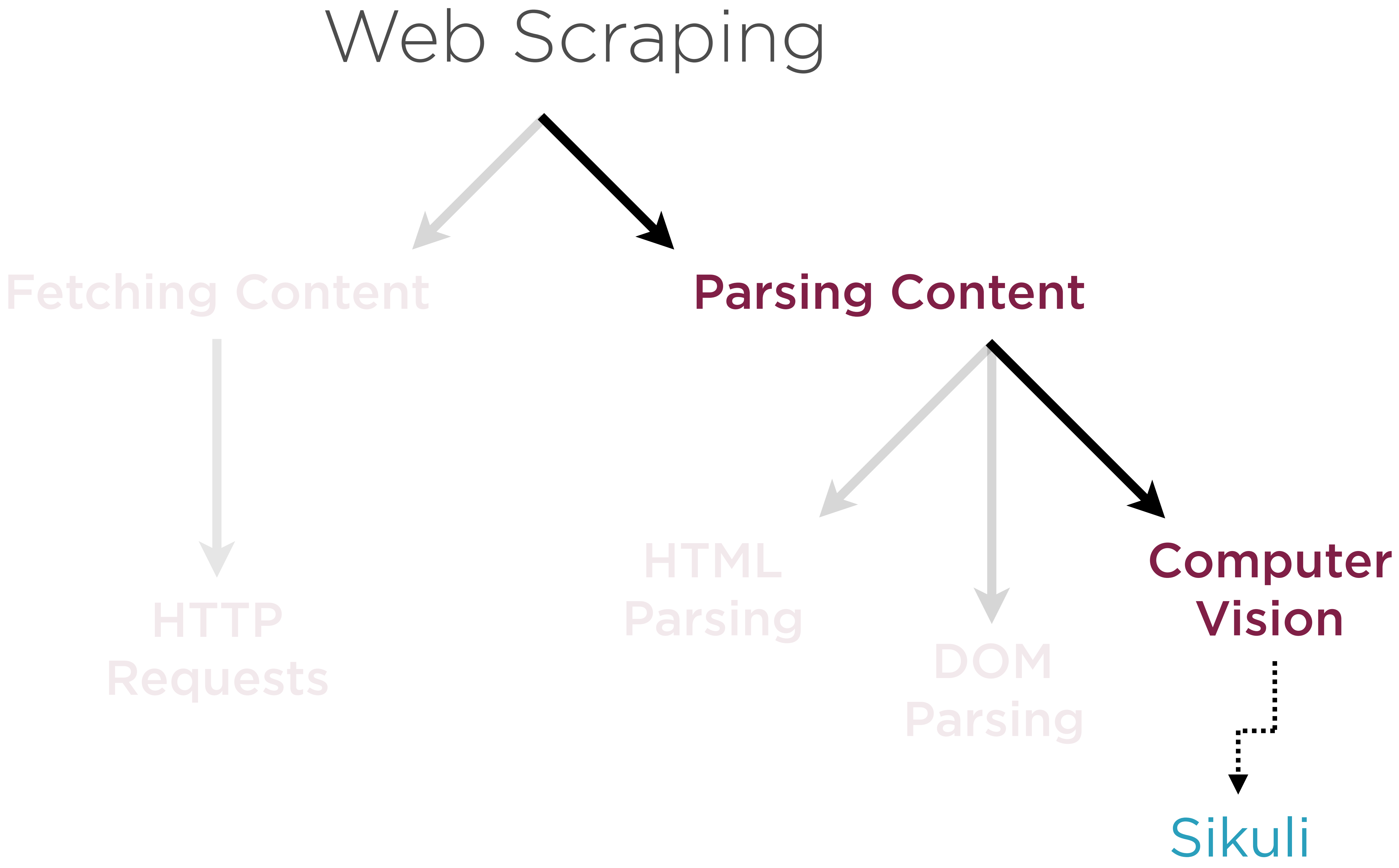
DOM Parsing

Computer Vision

Scrapy is a framework that combines all of this

Scrapy

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



Web Scraping

Fetching Content

Parsing Content

HTTP Requests

HTML Parsing

DOM Parsing

Computer Vision

Sikuli

Introducing Scrapy

Scrapy

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

https://doc.scrapy.org/en/latest/intro/overview.html

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

https://doc.scrapy.org/en/latest/intro/overview.html

Scraping vs. Crawling

Web Scraping	Web Crawling
Extract data directly from web sites	Download and index web sites
Data analysis and somewhat unsavory reputation	Performed by all search engines, associated with legitimate use
Specific - “scrape prices from Amazon”	General - “crawl sites linked off Amazon”
Small scale, results in specialized dataset	Large scale, results in document corpus

Scrapy

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

Framework vs. library: inversion of control

https://doc.scrapy.org/en/latest/intro/overview.html

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

https://doc.scrapy.org/en/latest/intro/overview.html

Scrapy

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

Inherently somewhat fragile, like regular expressions and other related tools

https://doc.scrapy.org/en/latest/intro/overview.html

Specific HTML elements are selected for processing using Selectors

Scrapy

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

https://doc.scrapy.org/en/latest/intro/overview.html

Scrapy supports selectors specified in CSS and XPath

Scrapy

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

https://doc.scrapy.org/en/latest/intro/overview.html

Selector

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

Scrapy Selectors

XPath

Select nodes in an XML (or HTML) document

CSS

Select HTML elements (usually to associate styles with them)

Scrapy selectors are built atop the lxml library

Demo

Installing Scrapy on our local machine

Exploring the Scrapy shell

Demo

Extracting content using CSS selectors

Demo

Extract content using XPath selectors

Summary

Scrapy as an application framework for crawling websites

Data extraction in a structured format

Interactively testing extraction using the Scrapy shell

Leveraging XPath and CSS selectors in the Scrapy shell