

Extracting Data from HTML with BeautifulSoup

GETTING STARTED WITH BEAUTIFUL SOUP



Janani Ravi

CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Understanding web scraping

Fetching web content via HTTP

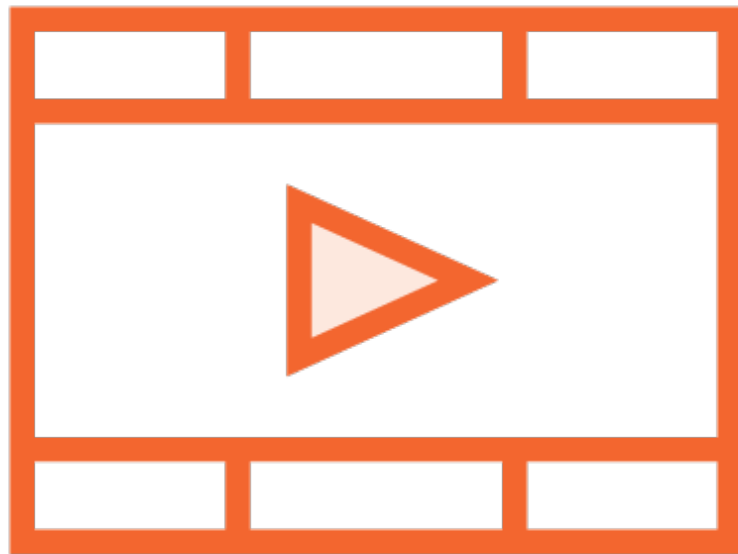
Regular expressions

Parsing HTML using regular expressions

Getting started with BeautifulSoup

Prerequisites and Course Outline

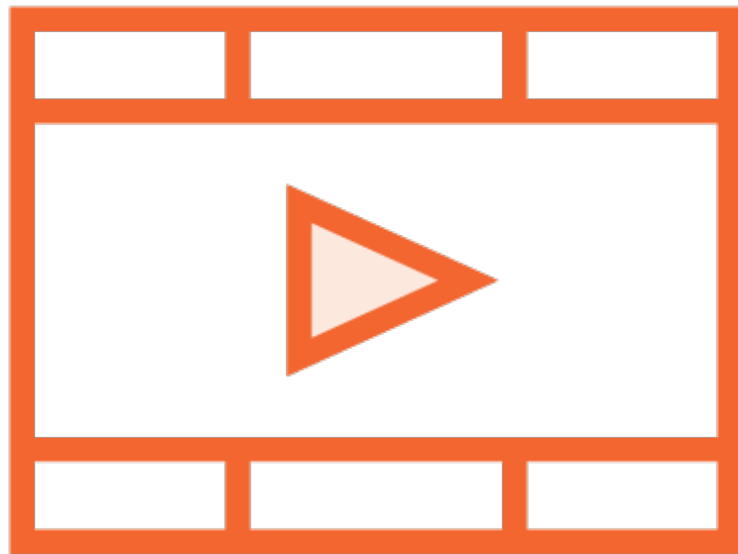
Prerequisites



Basic Python programming

**Basic knowledge of HTML, CSS
and web pages**

Prerequisite Courses



Python Fundamentals
Your First Day with HTML

Course Outline



Getting started with Beautiful Soup

Navigating the parse tree

Searching for elements in the parse tree

**Leveraging advanced features in
Beautiful Soup**

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 Scraping

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

Web Scraping

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

Web Scraping

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

Web Scraping

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

Web Scraping

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

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



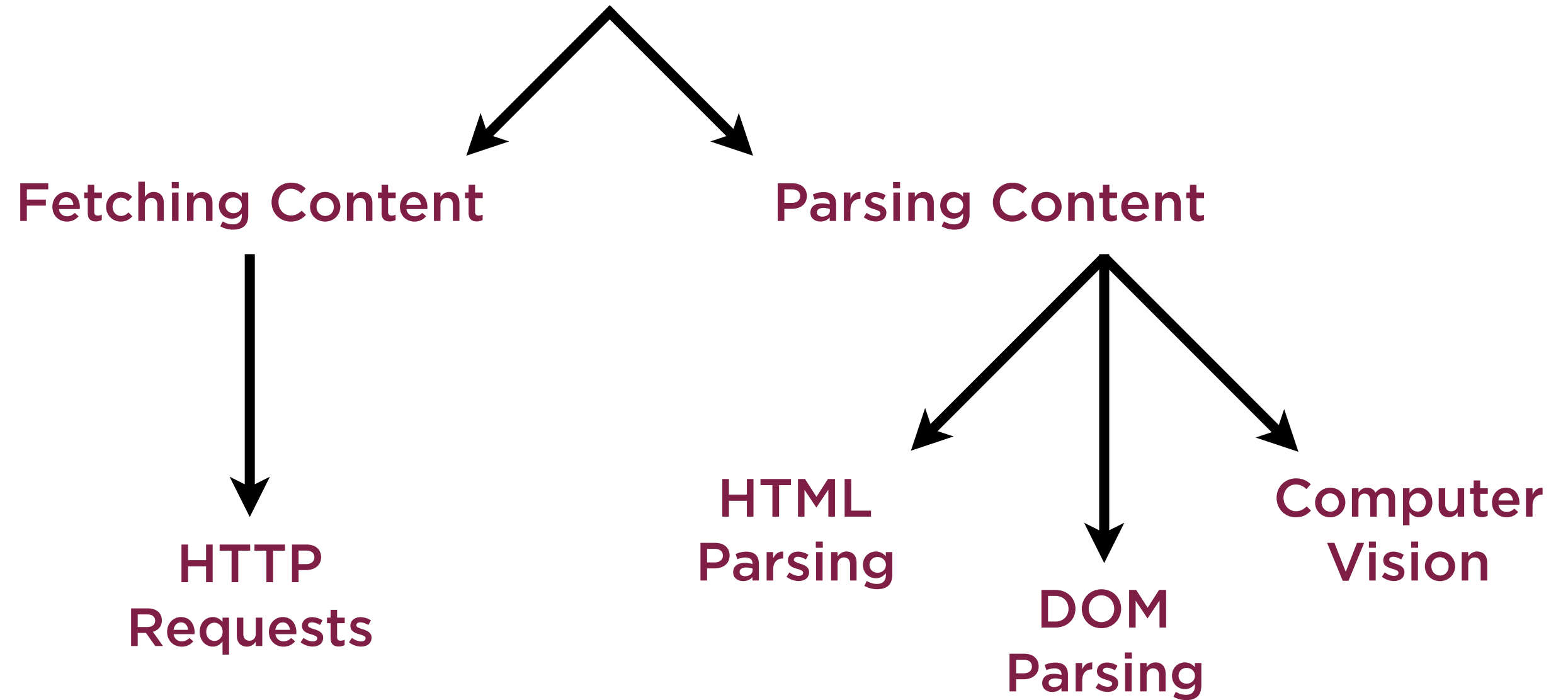
The same HTML markup can be accessed (**fetch**ed) via HTTP

Possesses an in-built hierarchical structure

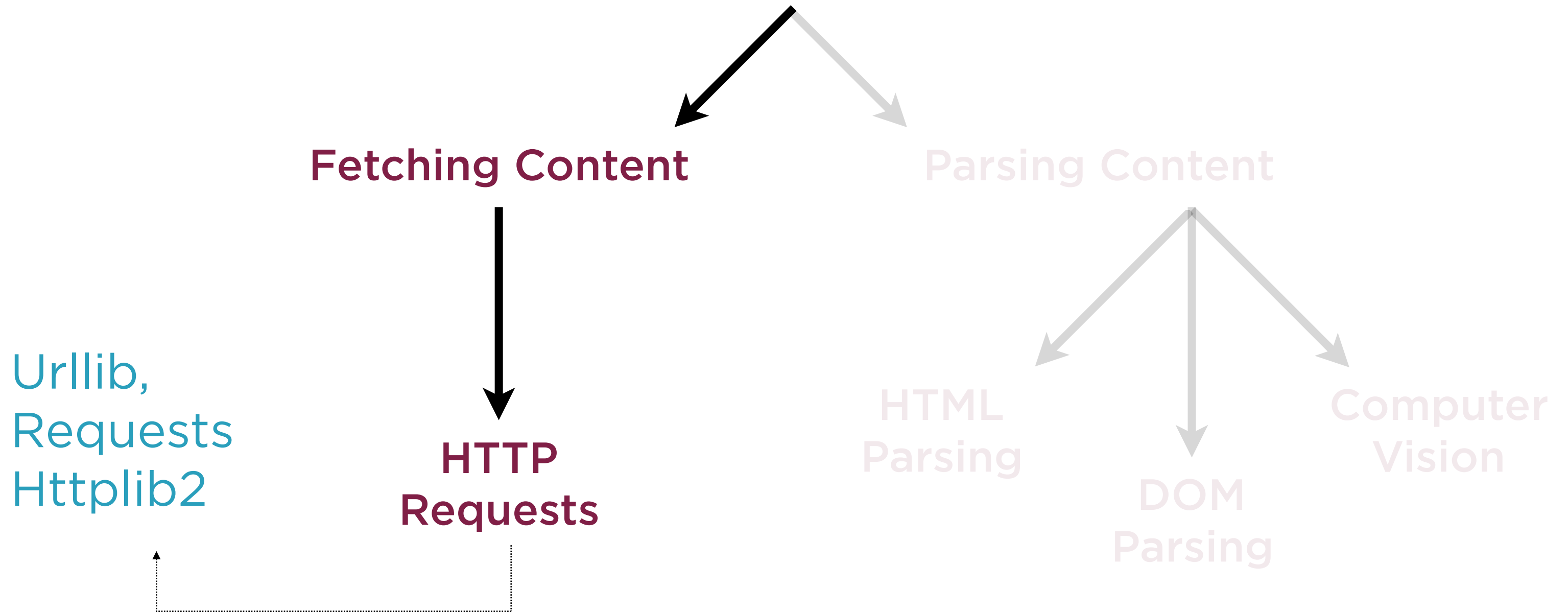
Parsers can exploit this structure to **extract** information

Fetching and Parsing Content

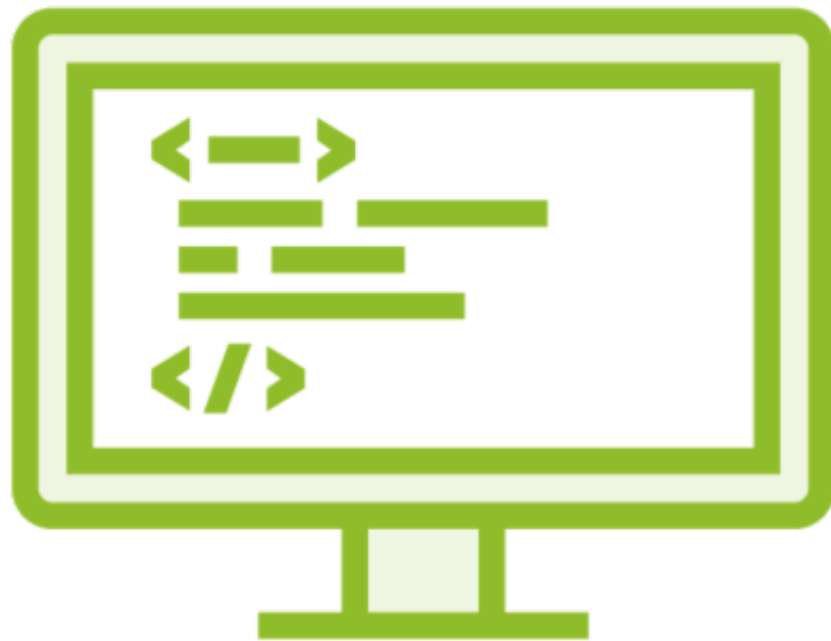
Web Scrapping



Web Scrapping



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



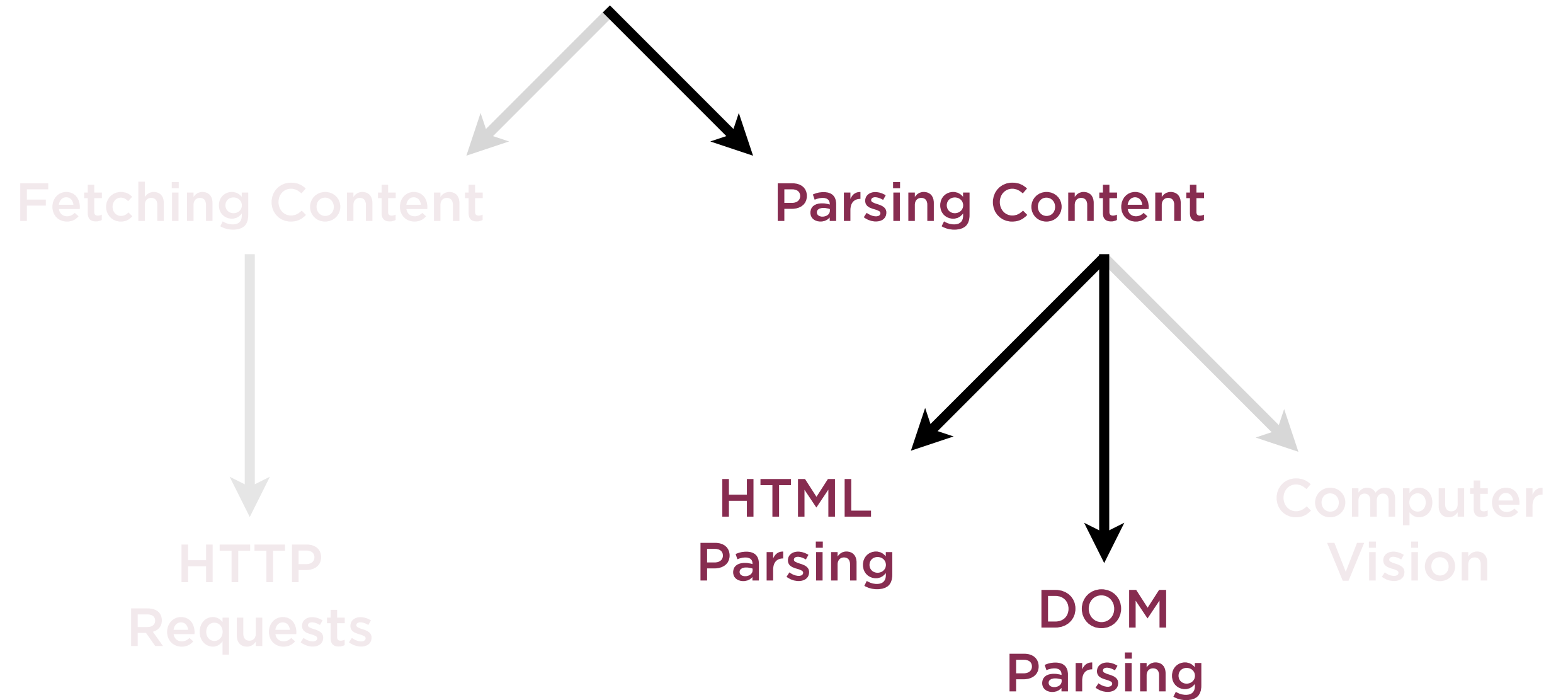
Command-line HTTP requests

- cURL

Python libraries for programmatic access

- Requests
- Httpplib2
- Urllib, Urllib2

Web Scrapping



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

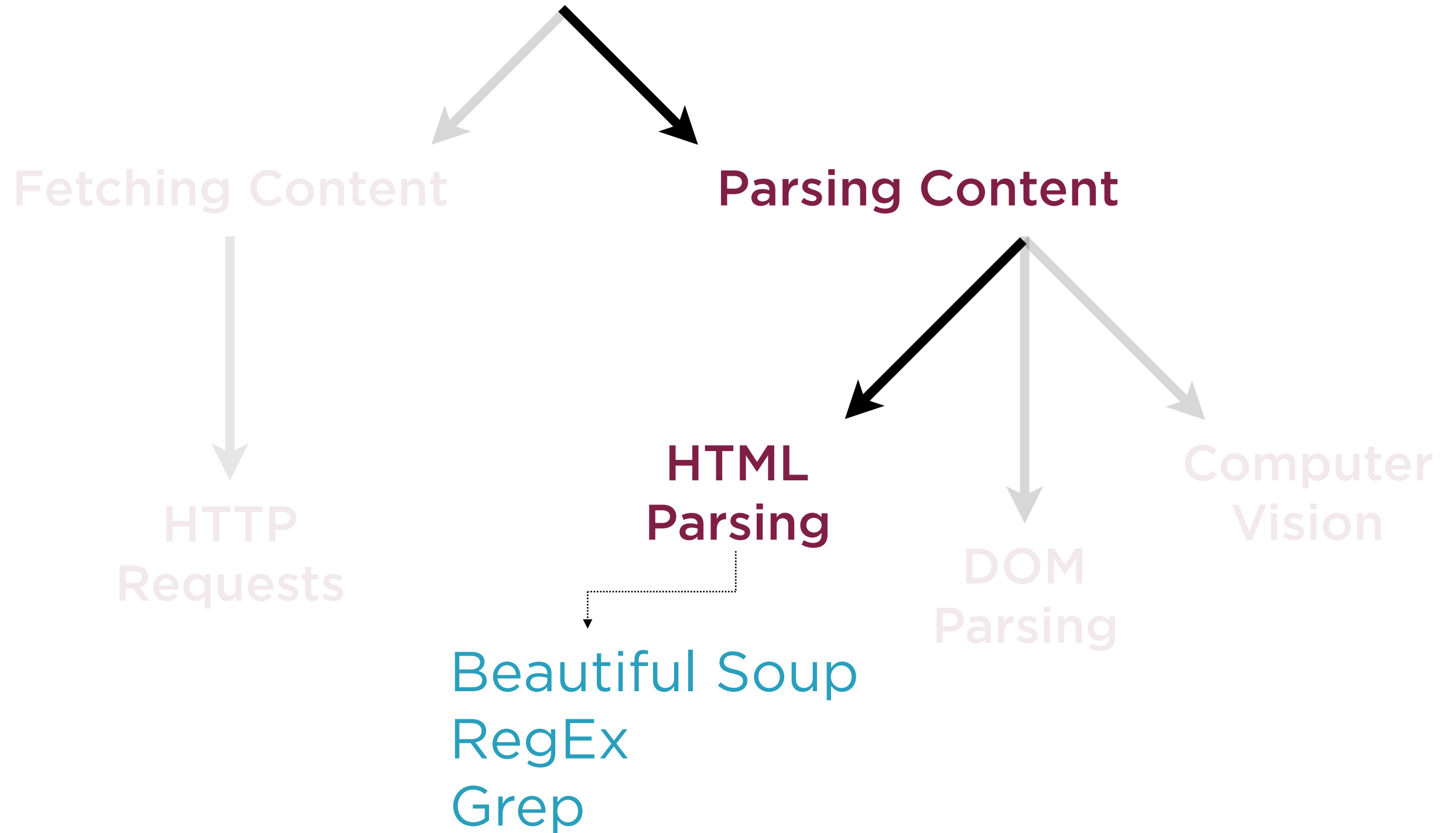
DOM Parsing

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

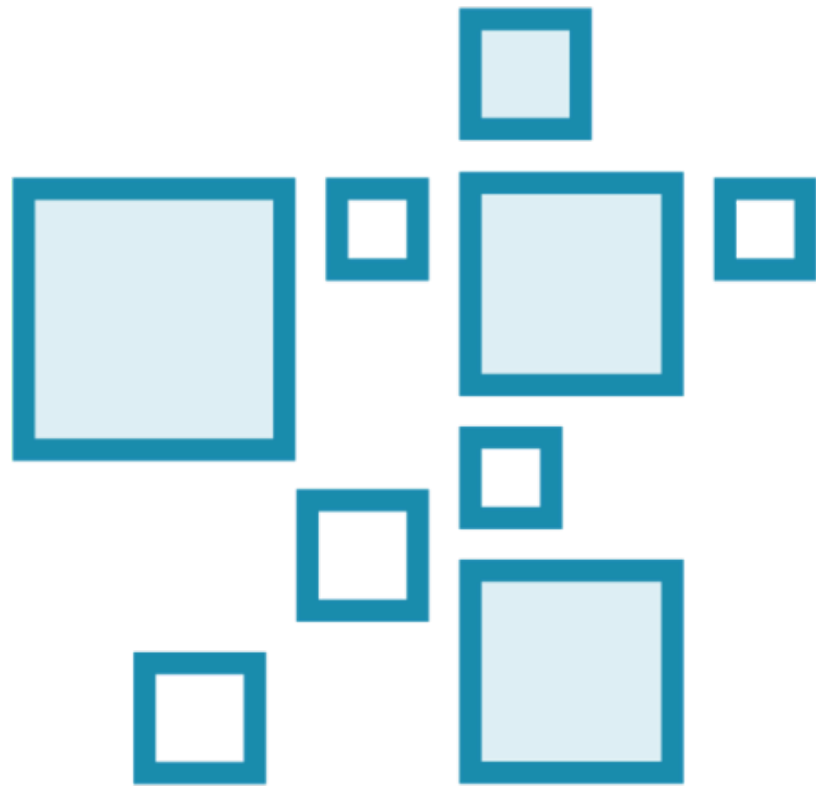
Web Scrapping



Regular Expressions

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

Regular Expressions



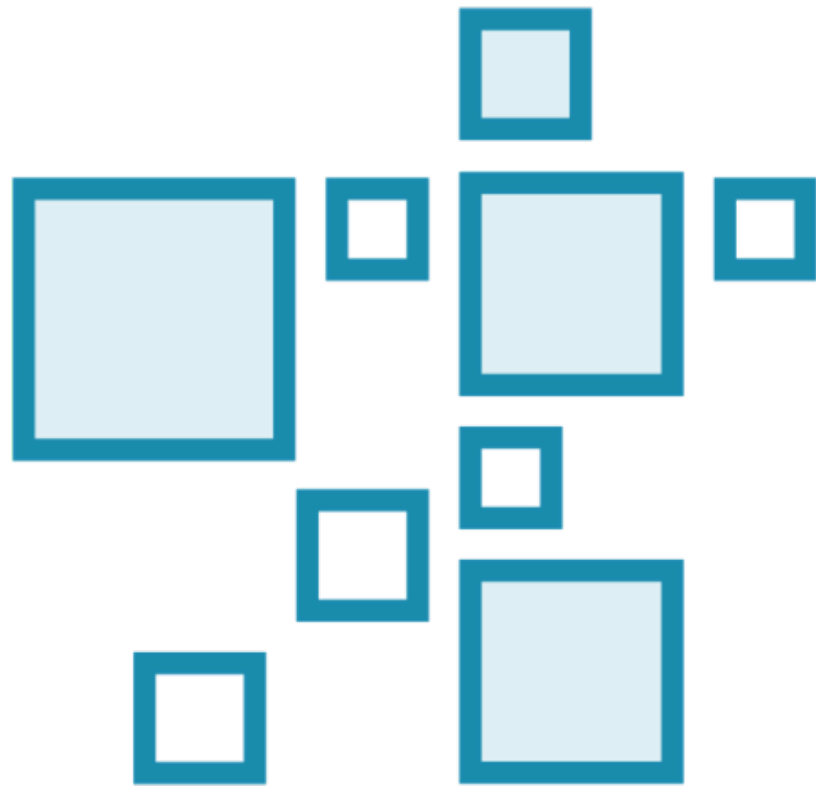
Regular expressions are based on text pattern matching

Very similar to grep and other Unix string utilities

Can be used from within virtually all programming languages

- Portable syntax

Regular Expressions



Regular expressions are a classic rule-based approach

- Local: Do not understand entire structure while matching
- Complex: Syntax is arcane and error-prone
- Fragile: Will not handle ill-formed HTML

Beautiful Soup

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

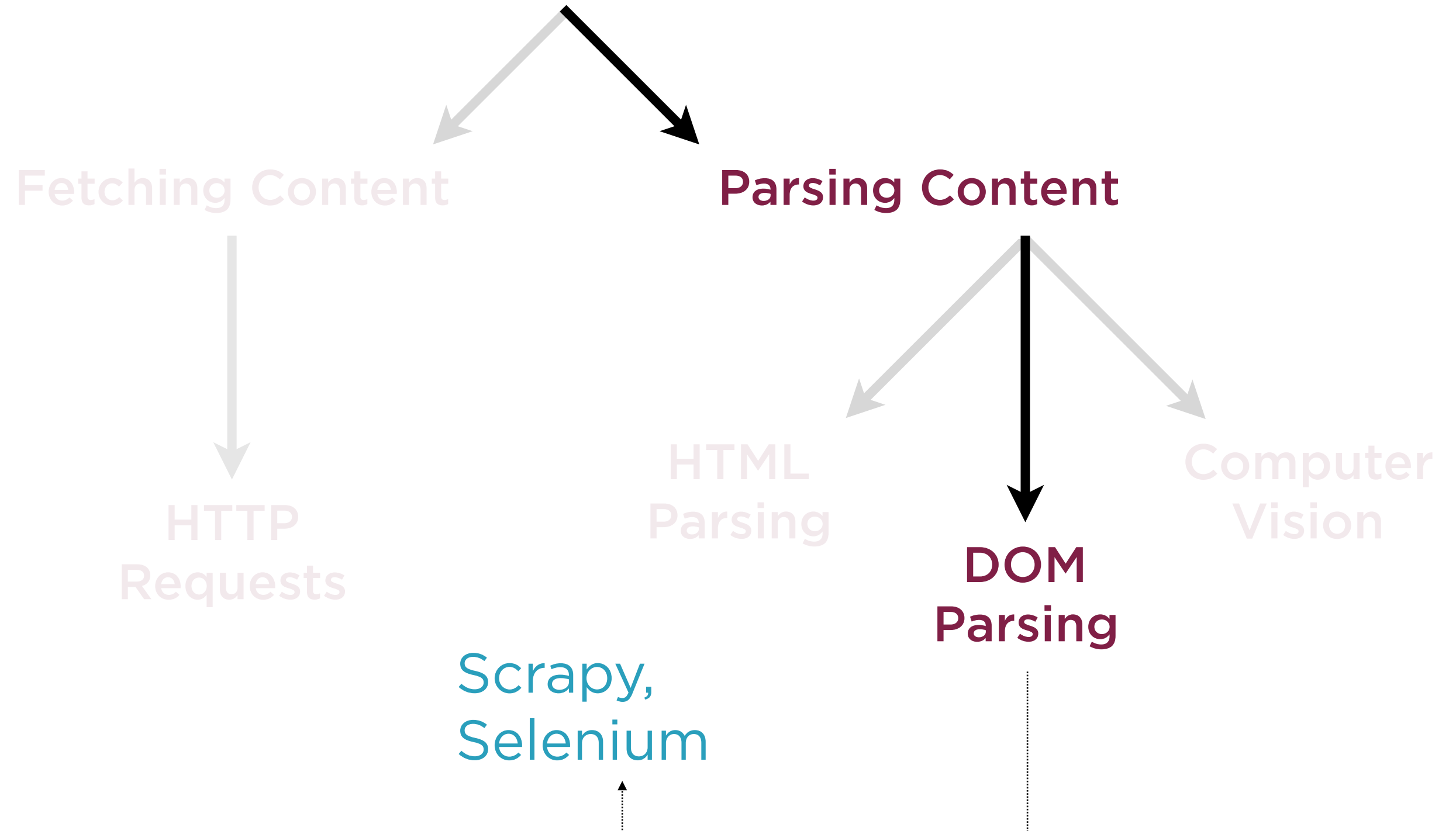
Beautiful Soup



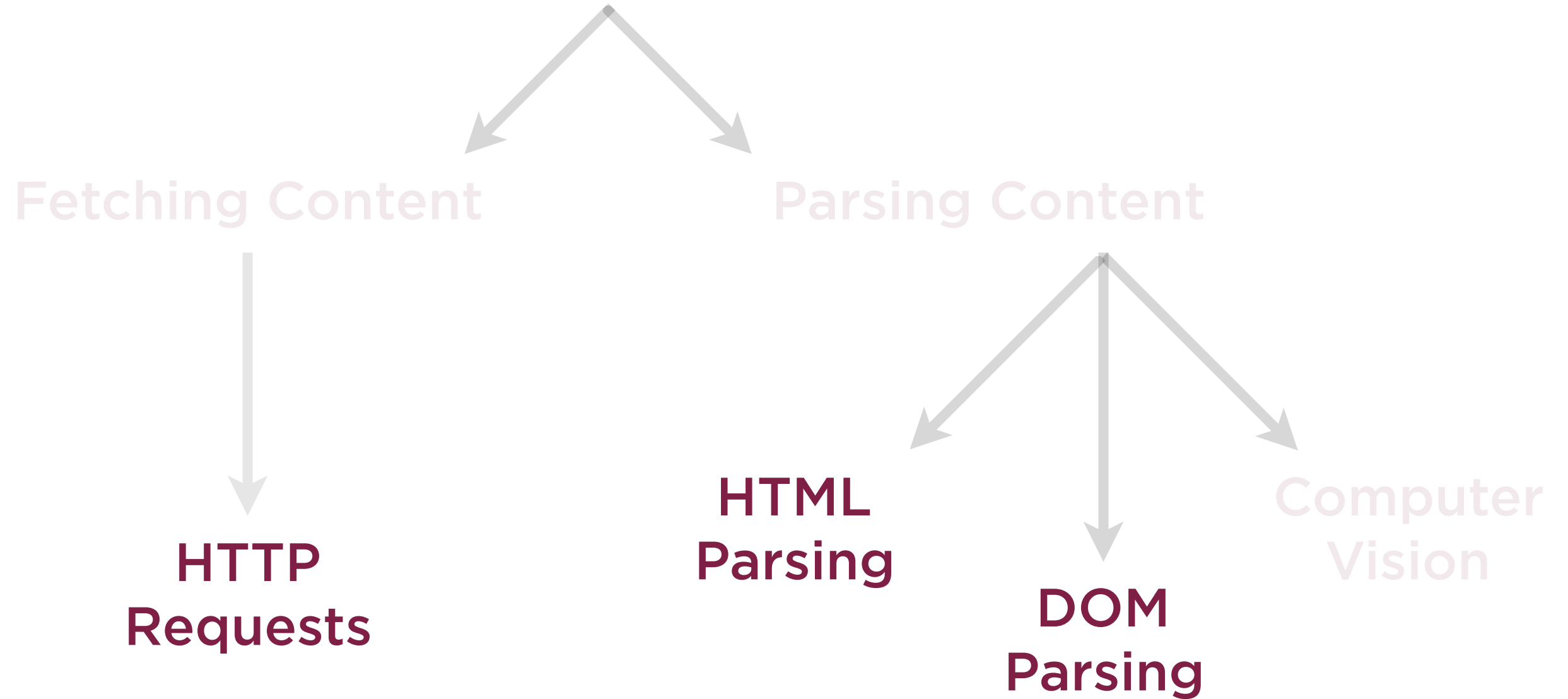
Mitigates weaknesses of regular expressions

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

Web Scrapping



Web Scrapping

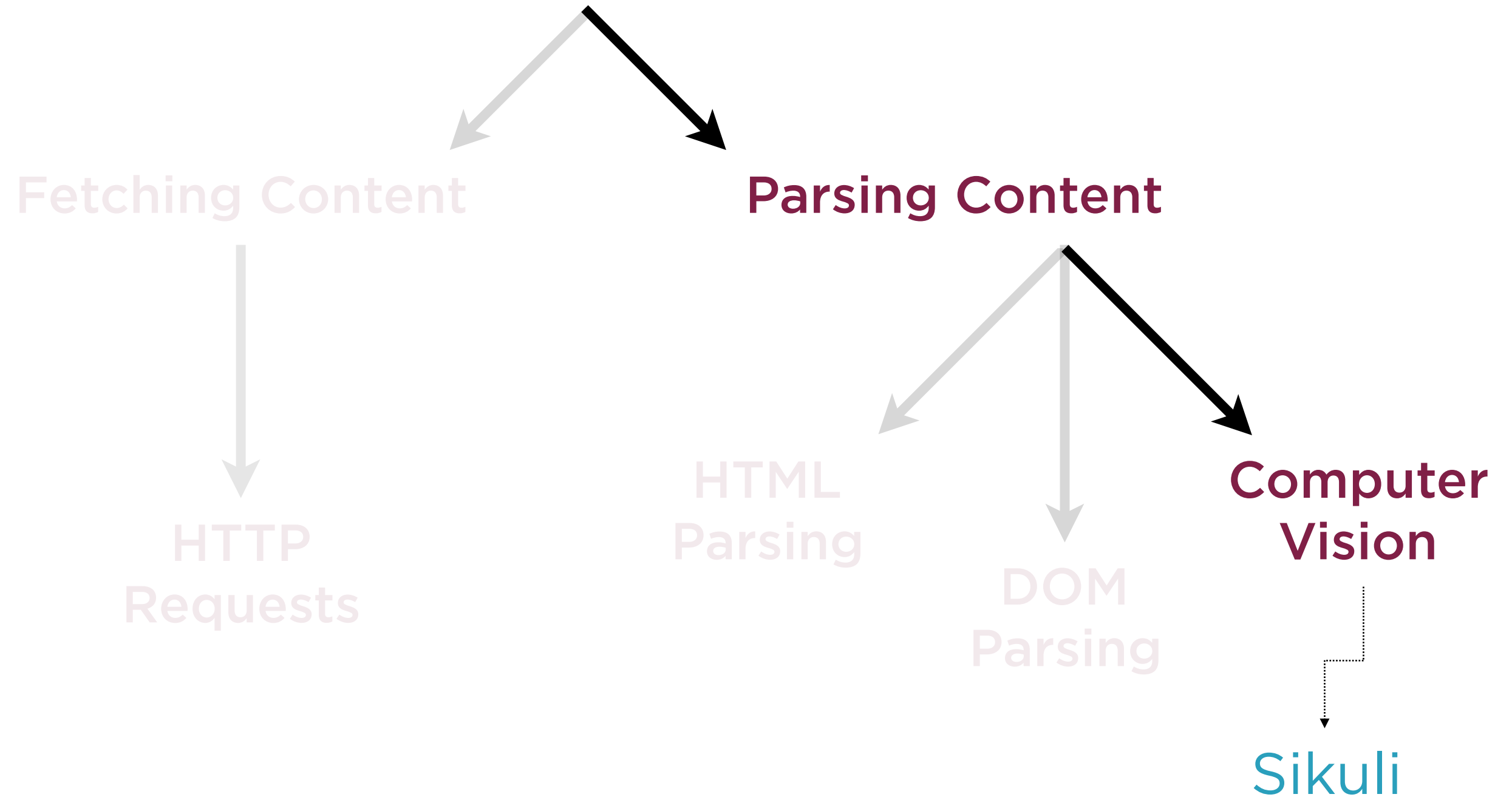


Scrapy is a framework that combines all of this

Scrapy

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

Web Scrapping



Demo

**Making HTTP GET requests using
different client libraries**

Demo

**Getting started with regular
expressions in Python**

Demo

Parsing web pages using regular expressions

Demo

Getting started with Beautiful Soup

Summary

Understanding web scraping

Fetching web content via HTTP

Regular expressions

Parsing HTML using regular expressions

Getting started with BeautifulSoup