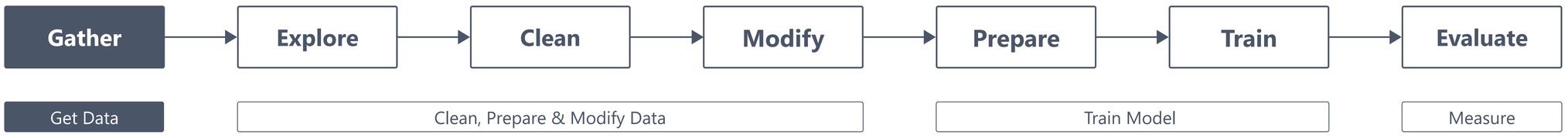
Web Scraping
Public DataSets (Kaggle, Azure, AWS, Google, ...)
Company Data Platform (Data lake)

AI Workflow

Data Gathering



WEB SCRAPING

WEB CRAWLING

A "Crawler" discovers the different websites and downloads them as HTML. It will then follow other pages (`<a href="#">`) and download these. Note that a robots.txt file located at the root explains to the Crawler which pages it can download.

Example *(javascript)*

```
import puppeteer from 'puppeteer';
const browser = await puppeteer.launch();
const page = await browser.newPage();
await page.goto('https://xaviergeerinck.com');
const html = await page.evaluate(() => document.documentElement.outerHTML);
console.log(html);
```

DATA PARSING & EXTRACTION

Once data is downloaded, it can be parsed and extracted. For this it's common to utilize a technique called "Regular Expressions" or XPaths.

Example *(javascript)*

```
// XPath
const xPath = '//html[1]/body[1]/div[1]';
document.evaluate(xPath, document, null, XPathResult.FIRST_ORDERED_NODE_TYPE, null).singleNodeValue;

// Regular Expression (RegEx)
const str = "Hello World!"
const re = new RegExp(/[A-Za-z]*/, 'g');
const matches = str.match(re);
--> ["Hello", "", "World", "", ""]
```

PUBLIC DATASETS

PROVIDER	URL
Kaggle	https://www.kaggle.com/datasets
Microsoft	https://azure.microsoft.com/en-us/services/open-datasets/catalog/
Google	https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset
Amazon	https://registry.opendata.aws/

COMPANY DATA PLATFORM

DESCRIPTION

A data platform is a centralized platform within an organization capturing a variety of different data sources and making them available to the different business units. This data platform takes care of tasks such ingestion, preparation and serving of the data to these different business units. With an ultimate goal of allowing business units to focus on their needs, without having to worry about cleaning and preparing their data in the correct format.

ARCHITECTURE EXAMPLE