XGEN R&D

Integrated Approach for Structured Internet Data Aggregation

Authors: Bogdan-Ioan Teglaș & Natanael Iosif Balogh
Students
UNIVERSITATEA TEHNICĂ DIN CLUJ-NAPOCA
Facultatea de Științe
Specializarea: Informatică

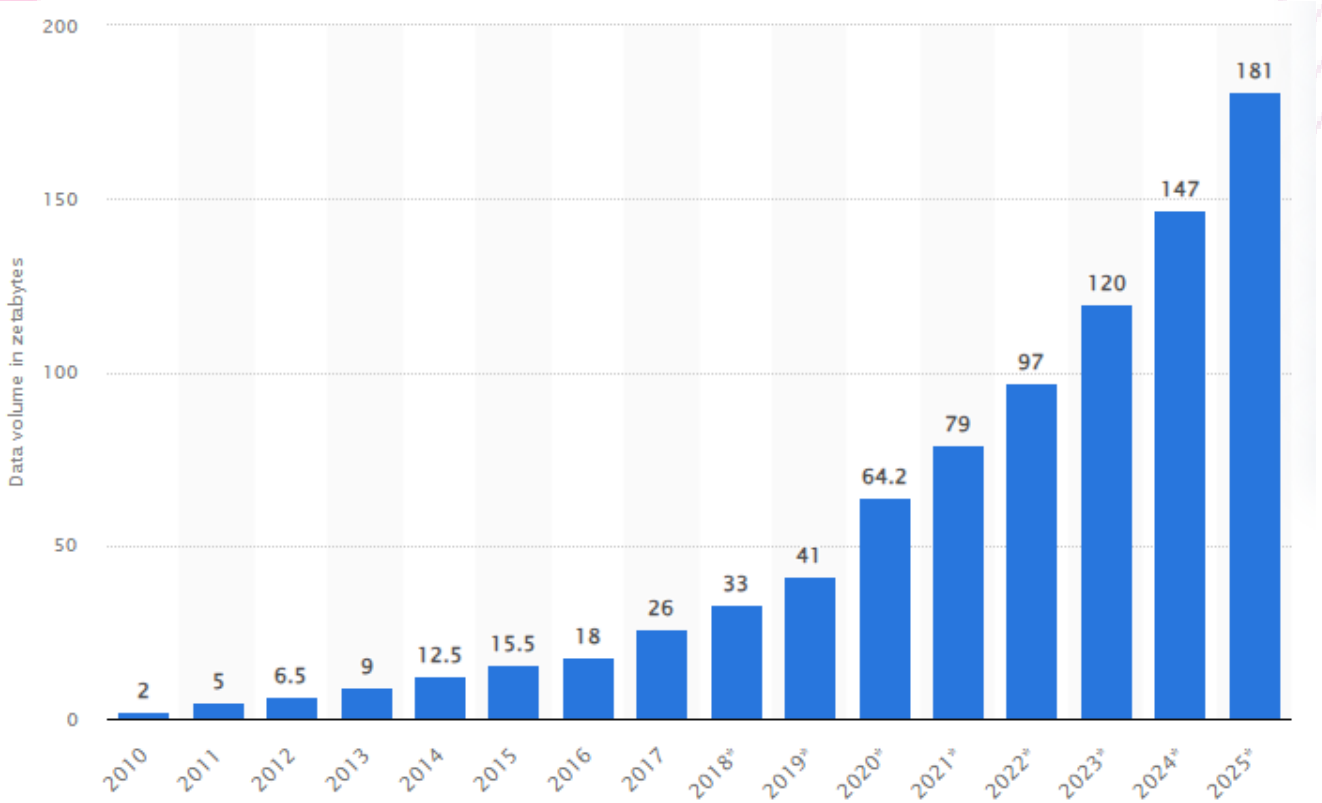
XGEN R&D

Introduction

- Big Data
- Applications
- Data to Knowledge

1. Introduction- Big Data

XGEN R&D



Digitalization

Active user contribution

Passive user contribution

1. Introduction- Applications

XGEN R&D

Data collected from the internet has multiple use cases in virtually any field of study or practice:

- Stock Market Analysis
- News Monitoring
- Social Media Sentiment Analysis
- Research Areas in Academia
- Hospitality Studies
- Socio-Economic Studies
- Business Strategies and Marketing
- Food Price Research
- Machine Learning Algorithm Training

1. Introduction- Data to Knowledge

Data
-raw-

Information
-processed-

Knowledge
-actionable-

2. Data collection

- Approaches
- Challenges

2. Data Collection- Approaches

ARTIFICIAL INTELLIGENCE

- INCREASINGLY POPULAR METHOD
- EMPLOYS MACHINE LEARNING TECHNIQUES
- PARTICULARLY ADVANTAGEOUS FOR HANDLING DIVERSE AND DYNAMIC DATASETS.
- ADDRESSES CHALLENGES POSED BY FREQUENTLY CHANGING WEB PAGE STRUCTURES.

RULE BASED PROGRAMMING

- UNIQUE CODE SCRIPTS ARE NEEDED FOR DIFFERENT TYPES OF WEB PAGES.
- SUBSTANTIAL CODING EFFORTS REQUIRED FOR PRECISE DATA TARGETING AND HANDLING EXCEPTIONS.
- PREFERRED METHOD IS USING APIS
- WEB SCRAPERS AND CRAWLERS ARE VALUABLE IN SCENARIOS WHERE THE ONLY AVAILABLE DATA SOURCE IS THE FRONT-END OF A WEB APPLICATION.

XGEN R&D

2. Data Collection- Challenges

Web Defensive Measures

Restrict Site Access

Page Loading Techniques

Website Navigation Obstacles

Problematic Data

2. Data Collection- Challenges

Other Challenges

Maintainability

Testing

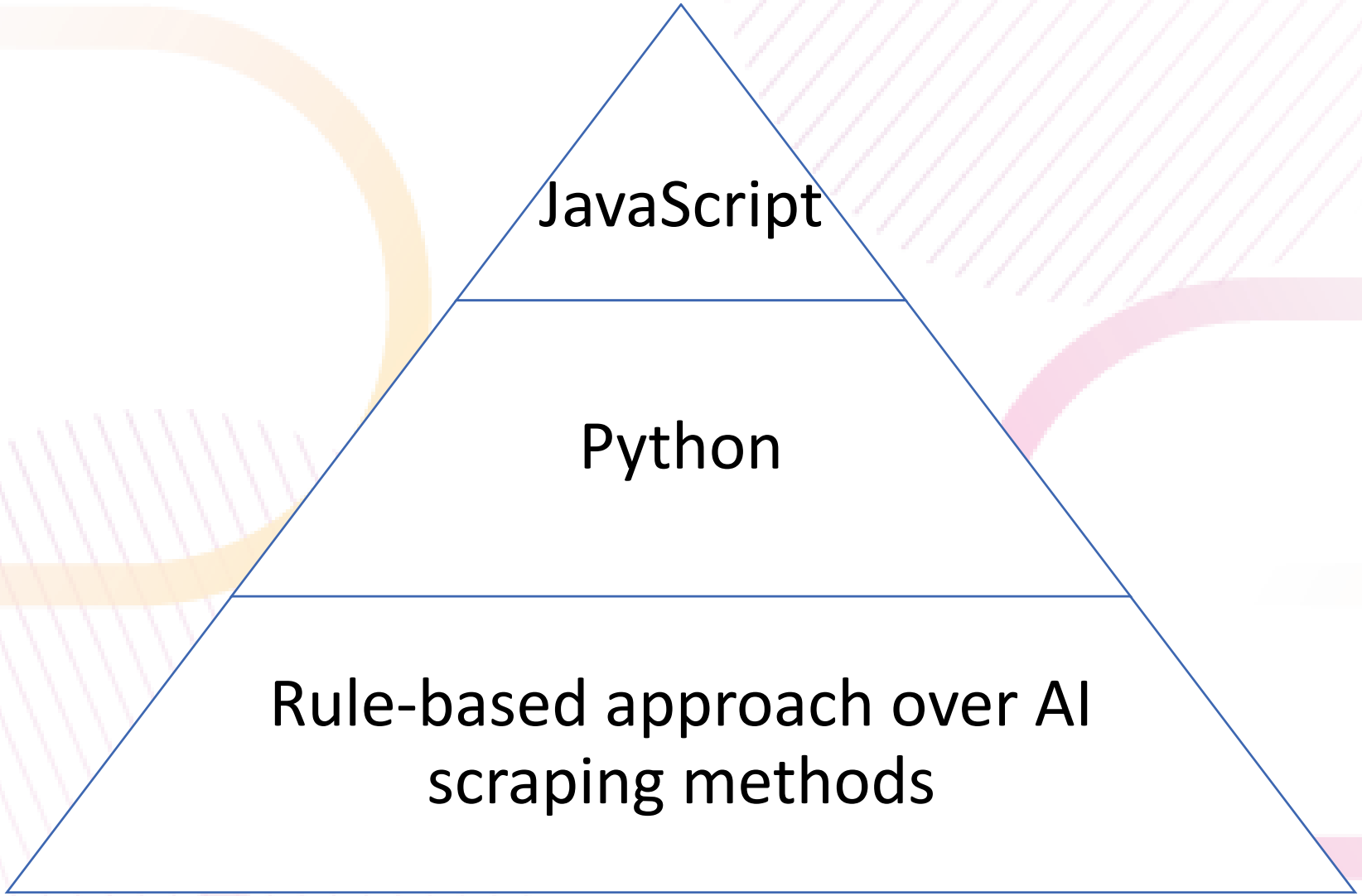
Customized Data Acquisition Objectives

XGEN R&D

3. Results

- Choosing the right tools
- Overcoming the specific challenges
- Demo Projects

3. Results- Choosing the right tools



3. Results- Overcoming the specific challenges 1

Consuming APIs	Strategies for Web Scraping Defense Mechanisms	Ensuring Data Accuracy, Consistency, and Completeness	Maintainability of Web Scraping Code
• Adherence to specified parameters for API requests • Parameters include request count, frequency, and proper use of API keys • Compliance with API documentation and real-time request simulation are crucial steps	• Replicating human interactions to mask automated patterns • Introducing delays and randomizing actions • Cookies management • Judicious use of proxies, considering IP address restrictions • Storing comprehensive information about errors and unexpected situations	• Developing meticulous checks for data consistency within each page • Continuous monitoring of web application behavior • Systematic checks to ensure the application functions as anticipated • Logging deviations from expected behavior for thorough analysis • Promptly addressing missing or incomplete data occurrences by adapting code	• Adherence to key programming principles and best practices • Robust modularization for easy implementation of minor changes • Designing independent modules to avoid interference • Meticulous function and class design for clarity and consistent outputs • Isolating configuration data for adaptability and scalability

3. Results- Overcoming the specific challenges 2

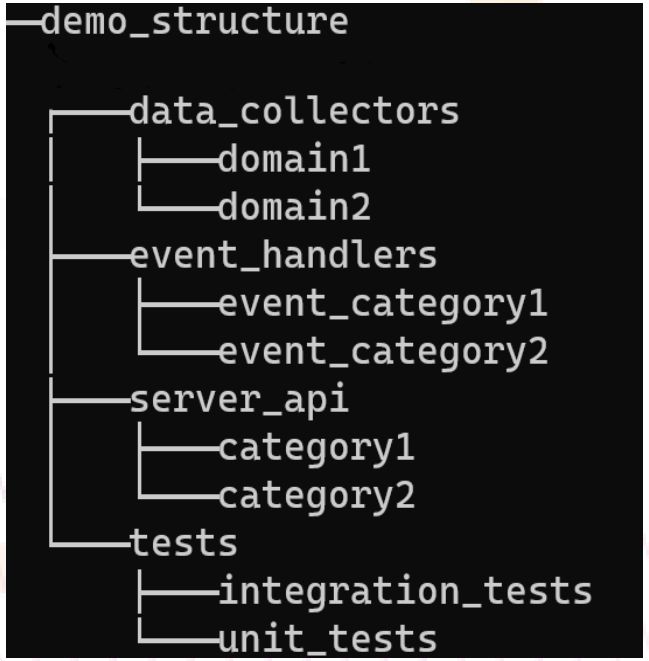
Alternative Strategy for Websites with Formidable Defenses

- Acquiring data through JavaScript implemented within a Google extension
- Leveraging a browser extension for controlled and adaptable data extraction
- JavaScript interacts with the Document Object Model (DOM) for nuanced retrieval

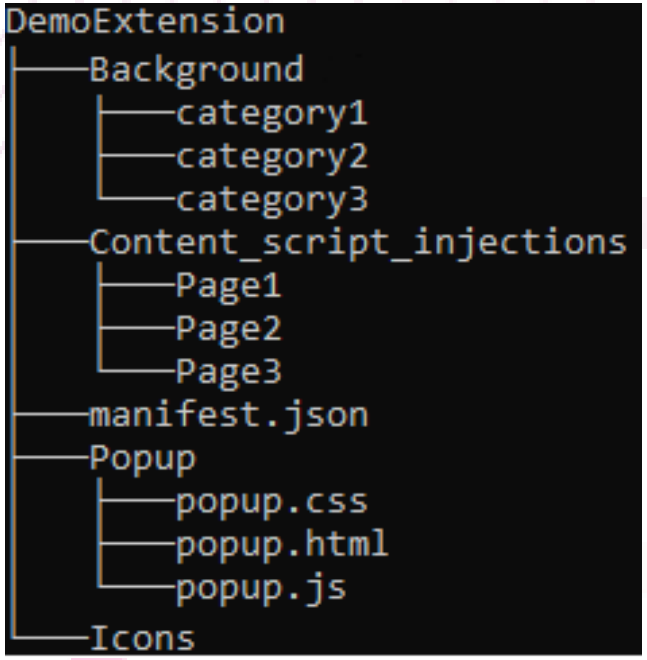
Vast Sources of Information and Too Much Redundant Data

- The volume of information poses a challenge for efficient data extraction
- Chrome extensions serve as a strategic solution to selectively acquire specific, relevant information
- User-centric, interactive nature of Chrome extensions allows tailored data extraction for efficiency

3. Results- Demo Projects



Automatic Approach: Python



Manual Approach: Chrome Extension

XGEN R&D

Thank you!