Data Ingestion and Big Data

Build a dataset from zero to solid

Mauro Pelucchi

mauro.pelucchi@gmail.com
https://github.com/mauropelucchi
https://www.linkedin.com/in/mauropelucchi/

EUROPYTHON
2021 Jul 26 - Aug 1 Online

Description

Web scraping, crawling and API are the first step to retrieve information to use for analysis and to start a new business.

In this tutorial I'll show you how to use python to set up scraping and crawling processes, how to simulate users navigation and browser behavior with a ghost browser and how to hook up and use data APIs. I will also try to explain the technical and ethical aspects that we have to consider when we approach these kinds of challenges..

You are provided a git repository (https://github.com/mauropelucchi/europython2021) with:

- Jupyter notebooks

- Slides

During this training we will use **Google Colab**

Mauro Pelucchi

I am a senior data scientist and big data engineer responsible for the design of the "**Real-Time Labour Market Information System on Skill Requirements**" for CEDEFOP

I currently works for Emsi | burning glass

My main tasks are related to machine learning modelling, labour market analyses, and the design of big data pipelines to process large datasets of online job vacancies.

In collaboration with the **University of Milano-Bicocca**, I took part in many research projects related to the labour market intelligence systems.

I collaborate with the **University of Milano-Bicocca** as a lecturer at the Master Business Intelligence and Big Data Analytics and with the **University of Bergamo** as a lecturer in Computer Engineering.

Agenda

- Data ingestion patterns
- Scraping vs crawling
- How to build a dataset from web data
- Introduction to phantom browsers
- CSS Selector
- Scraping hands-on with Selenium
- Ethics

Data ingestion pattern

Big Data was initially created to serve as storage for all raw data, with access to any kind of historical information always available for corporations.

Over time, **Big Data distributions became more and more effective**, not only at ingesting and scaling the data, but also providing **data processing capabilities for stored and real-time data**.

Ready for a Challenge?

What is **web scraping**?

A. Cleaning fake news from social media

B. Extracting structured content from a web page through a programme

C. Finding offers online

D. Cleaning out your garage

Scarping vs Crawling

A **web crawler** (spider) is an Internet bot that **systematically** browses the World Wide Web, typically for the **purpose of Web indexing** (web spidering). Web search engines and some other sites use Web crawling or spidering software to update their web content or indices of others sites' web content.

The crawler navigate through a given website and typically **store all data**.

A **web scraping** tool is a technology solution to **extract data** from web sites, in a **quick**, **efficient** and **automated** manner, offering data in a **more structured** and easier to use format.

Web scrapers simulate human exploration of the World Wide Web by either implementing low-level hypertext transfer protocol or embedding suitable Web browsers.

Scarping vs Crawling

Web Scraping, main tasks

- Internet: most data is unstructured
  - Identifying pages
  - Understanding their structure
  - Navigating (e.g. browsing a list of results)
  - Structuring the data collection
    - Associating a structured field with a portion of the page
    - Building a structured dataset

Why?

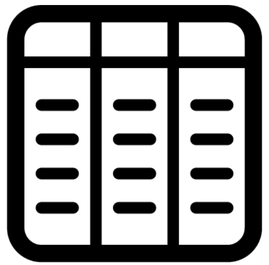
- For research (academic, marketing, ...)
  - Data are not available in structured format, no database exists
  - Multiple sources
- For business/eCommerce
  - Market Analysis
  - Product & Price Comparison
  - Online reputation
- For Marketing/CRM
  - What do they say about me (public opinion)?
  - Where is the market moving?
  - What do the media think (e.g. scraping newspaper articles)?
  - Better target ads for customers
  - Detect Fraudulent Reviews
- ...

Tools

- Visual crawling and scraping tools
  - Import.io www.import.io
  - Mozenda https://www.mozenda.com
  - Octoparse www.octoparse.com
- Framework
  - Nutch and Solr http://nutch.apache.org/
  - HTTrack www.httrack.com
  - Stormcrawler
- Custom code
  - JSOUP www.jsoup.org (Java), SCRAPY https://scrapy.org (Python)
  - **BeautifulSoup https://pypi.org/project/beautifulsoup4/ (Python)**
  - **Selenium https://www.selenium.dev/**

https://github.com/mauropelucchi/europython2021

INDIEGOGO



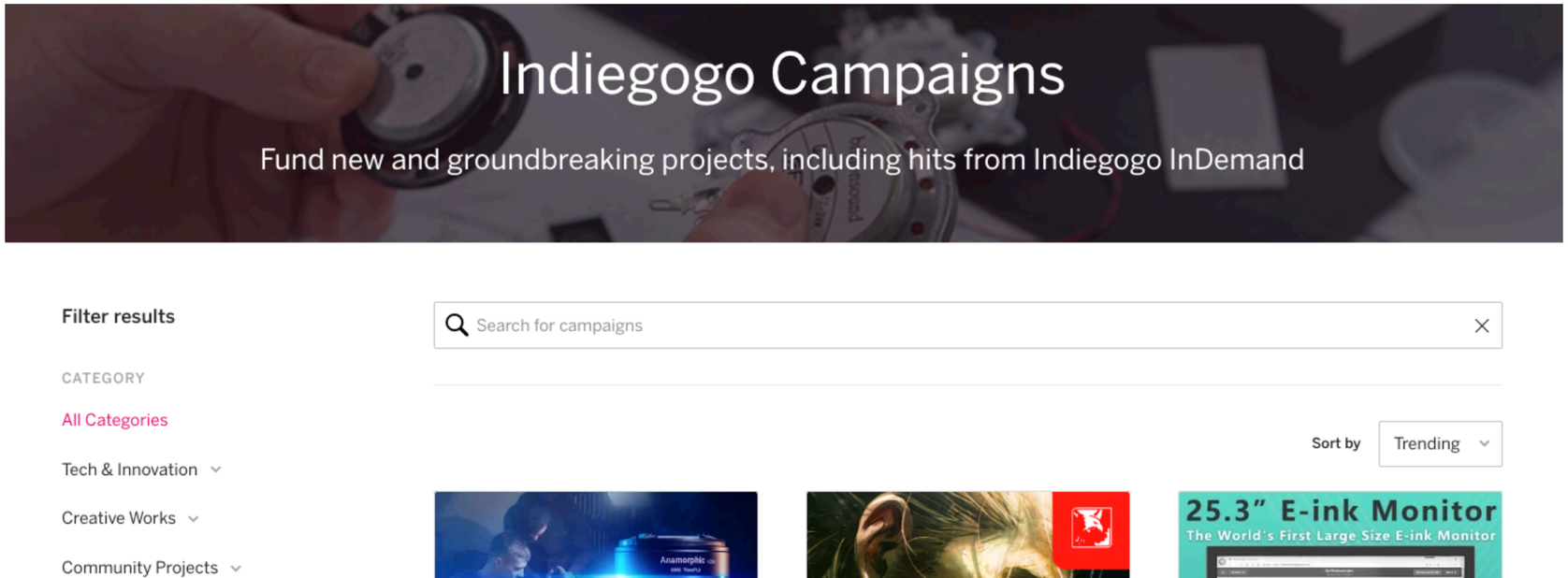
python + colab

The browser

The browser = The best way to **surf** the Internet

Open a web page

—> https://www.indiegogo.com/explore/all?project_type=campaign&project_timing=all&sort=trending



Web browsers use HTML (**HyperText Markup Language**) to display web pages.

HTML

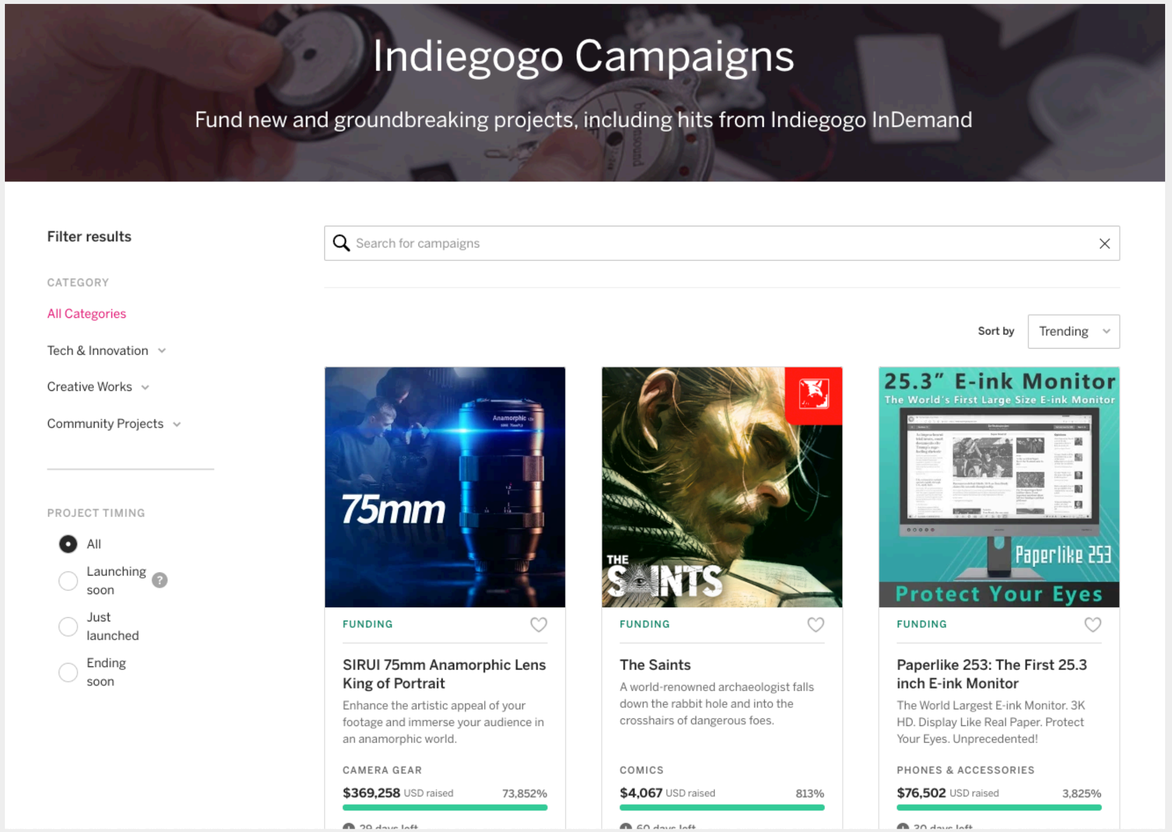
- Is composed of elements: **TAG**
- Each element has a start tag **<element>** and a closing tag **</element>**
- Each element can have an **ID**, which should be unique per page

<element id="ads"> </element>

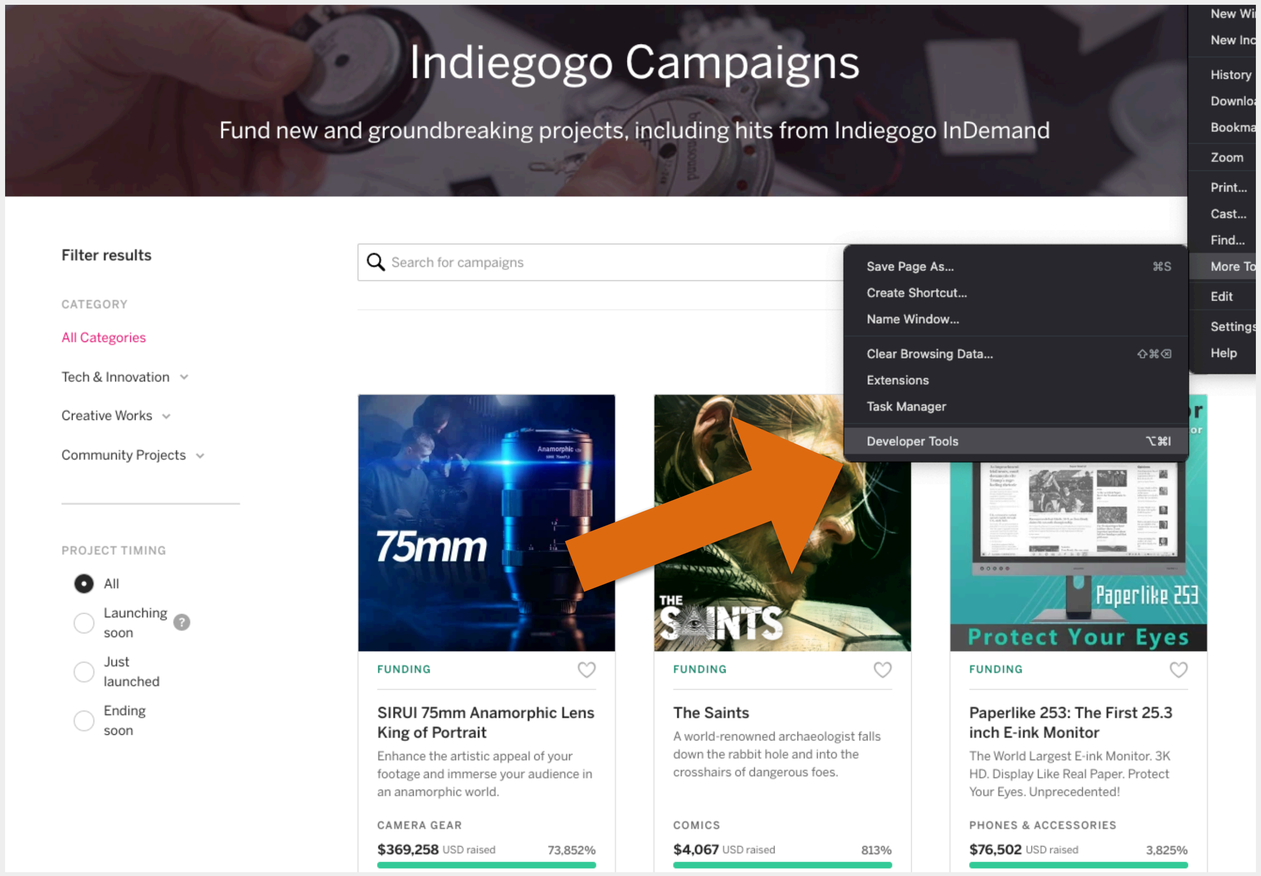
- **Classes**: used to categorise elements and apply styles. There may be many elements with the class “my_class".

<element class = “my_class”>

</ element>



How can I view the structure of the web page?

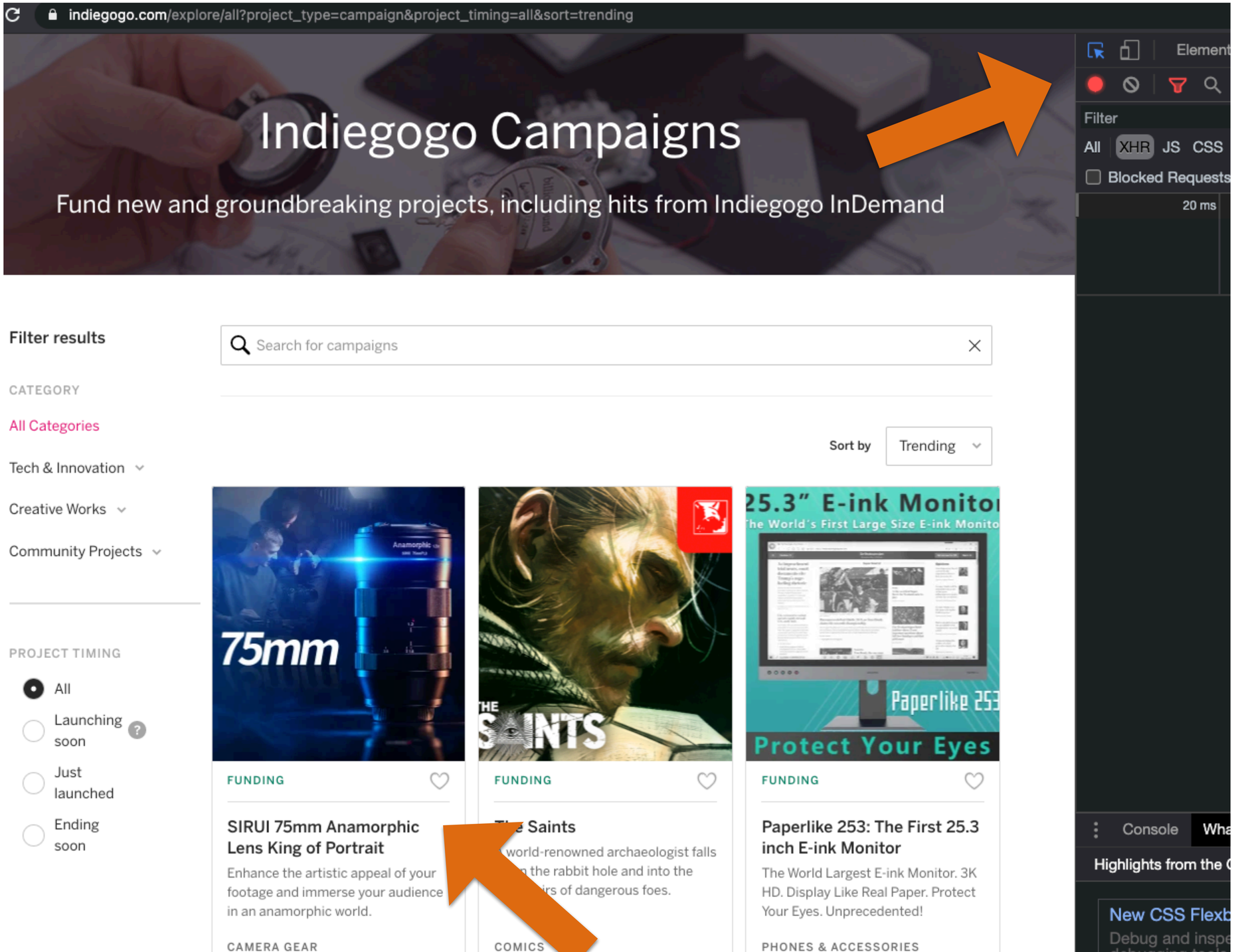


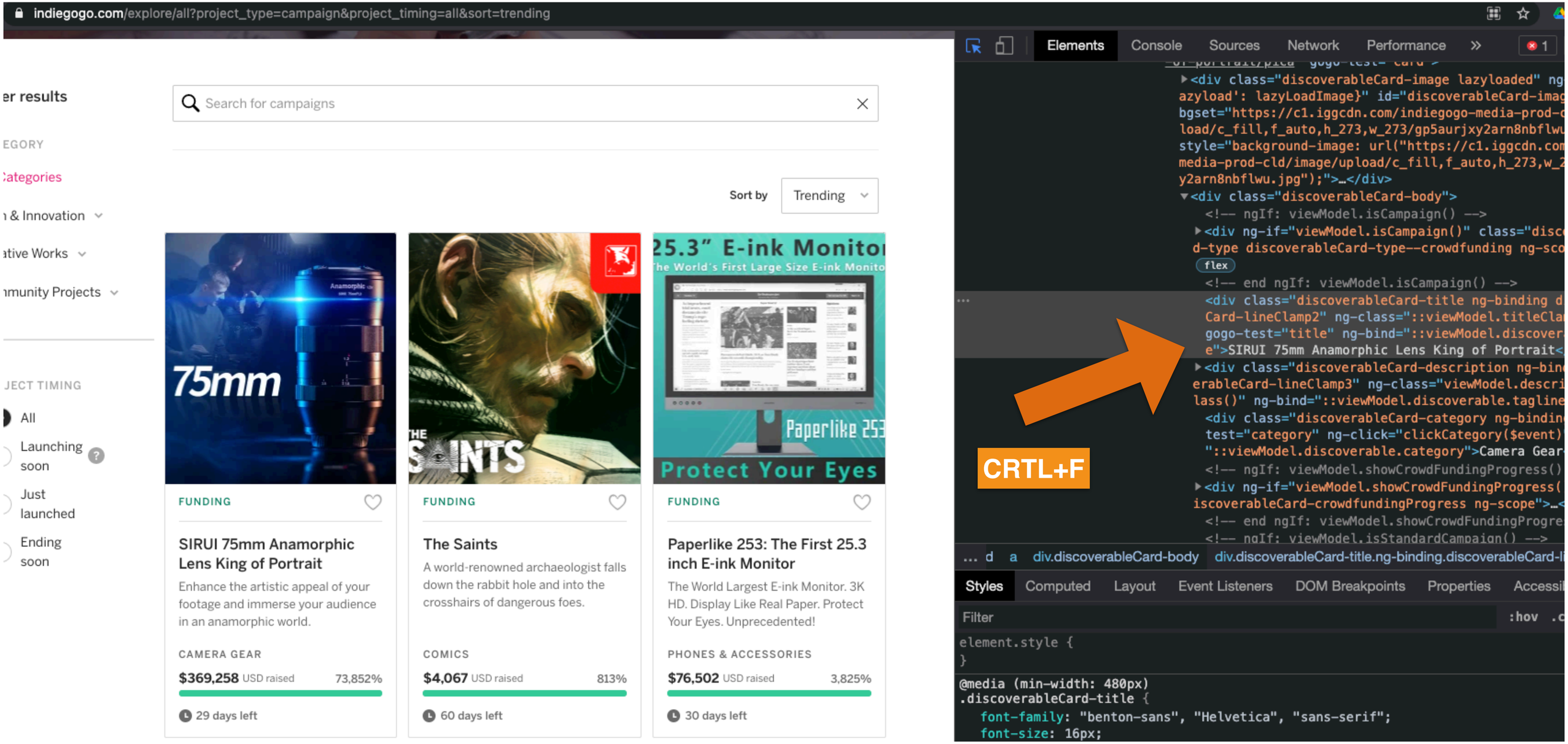
Developers Tools

Google Chrome

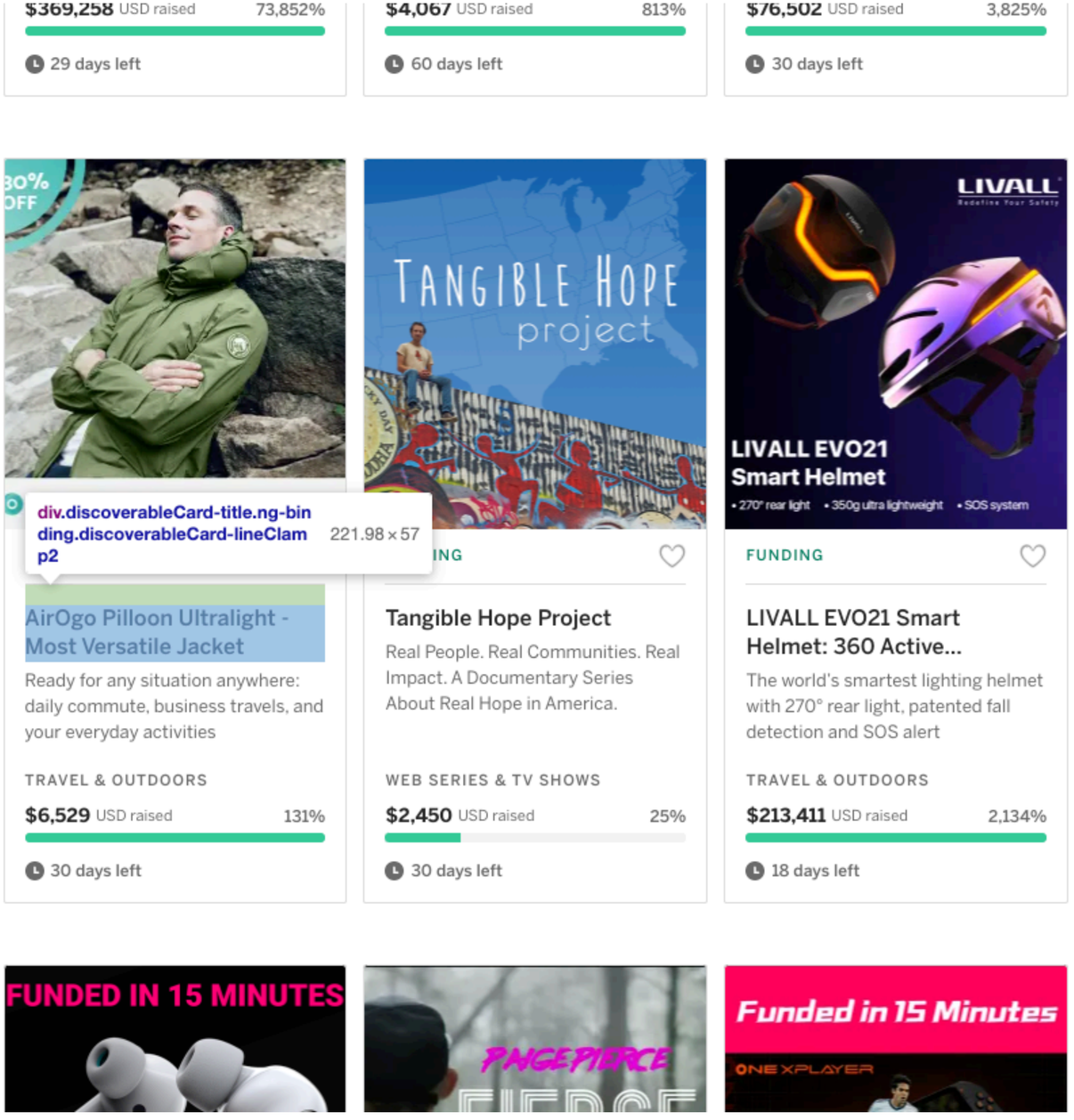
Menu> More tools> Developers tools

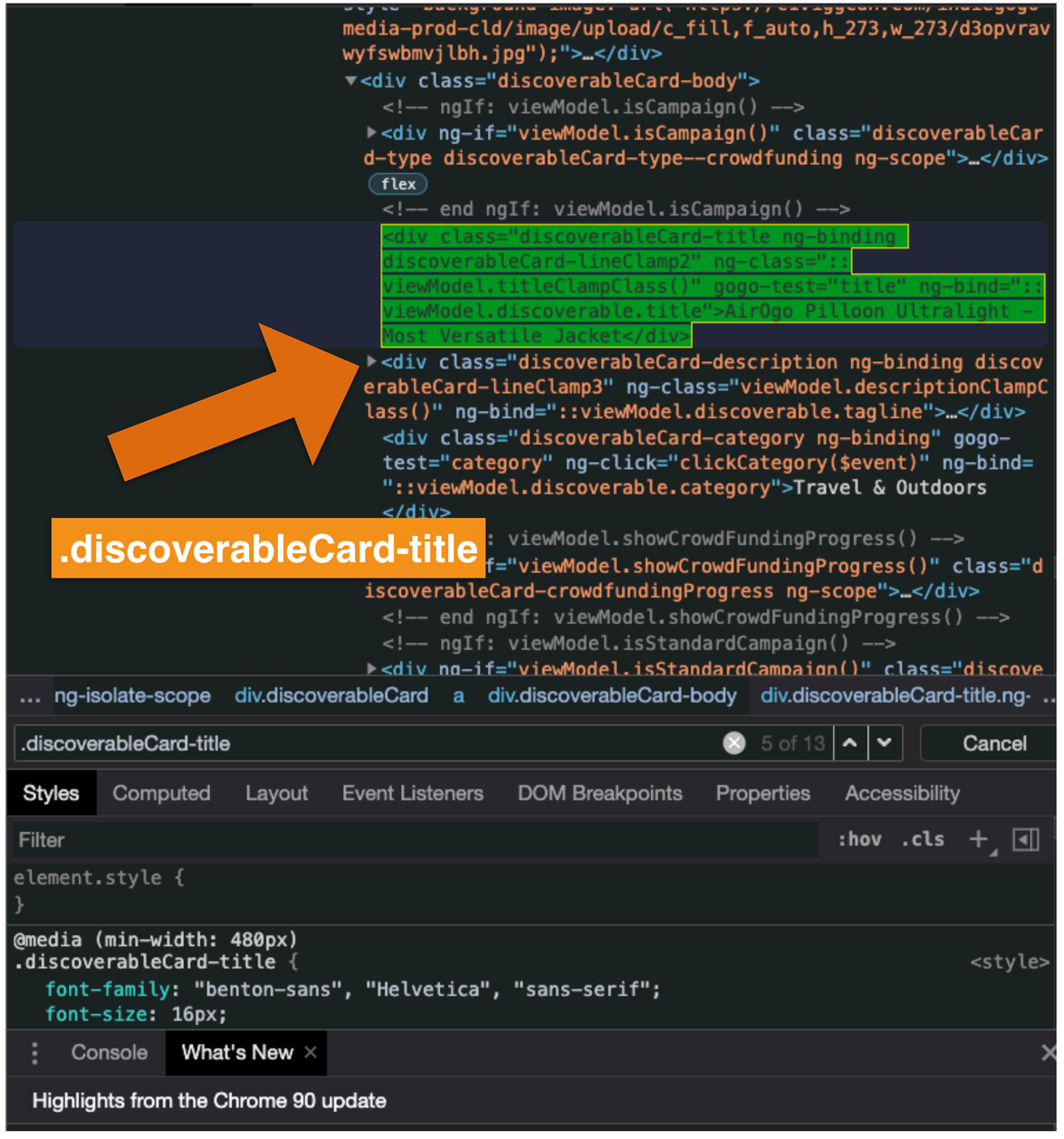

Start
End
Google
Attribute
Text


indiegogo.com/explore/all?project_type=campaign&project_timing=all&sort=trending
Indiegogo Campaigns
Fund new and groundbreaking projects, including hits from Indiegogo InDemand
Filter results
Search for campaigns
CATEGORY
All Categories
Tech & Innovation
Creative Works
Community Projects
PROJECT TIMING
All
Launching soon
Just launched
Ending soon
Sort by
Trending
75mm
FUNDING
SIRUI 75mm Anamorphic Lens King of Portrait
Enhance the artistic appeal of your footage and immerse your audience in an anamorphic world.
CAMERA GEAR
THE SAINTS
FUNDING
COMICS
25.3" E-ink Monitor
The World's First Large Size E-ink Monitor
Paperlike 253
Protect Your Eyes
FUNDING
Paperlike 253: The First 25.3 inch E-ink Monitor
The World Largest E-ink Monitor. 3K HD. Display Like Real Paper. Protect Your Eyes. Unprecedented!
PHONES & ACCESSORIES
Element
Filter
All
XHR
JS
CSS
Blocked Requests
20 ms
Console
Highlights from the
New CSS Flexb
Debug and inspe


indiegogo.com/explore/all?project_type=campaign&project_timing=all&sort=trending
er results
Search for campaigns
EGORY
Categories
& Innovation
ative Works
mmunity Projects
JECT TIMING
All
Launching soon
Just launched
Ending soon
Sort by
Trending
75mm
Anamorphic
THE SAINTS
25.3" E-ink Monitor
The World's First Large Size E-ink Monitor
Paperlike 253
Protect Your Eyes
FUNDING
SIRUI 75mm Anamorphic Lens King of Portrait
Enhance the artistic appeal of your footage and immerse your audience in an anamorphic world.
CAMERA GEAR
$369,258 USD raised
73,852%
29 days left
FUNDING
The Saints
A world-renowned archaeologist falls down the rabbit hole and into the crosshairs of dangerous foes.
COMICS
$4,067 USD raised
813%
60 days left
FUNDING
Paperlike 253: The First 25.3 inch E-ink Monitor
The World Largest E-ink Monitor. 3K HD. Display Like Real Paper. Protect Your Eyes. Unprecedented!
PHONES & ACCESSORIES
$76,502 USD raised
3,825%
30 days left
Elements
Console
Sources
Network
Performance
<div class="discoverableCard-image lazyloaded" ng
azyload': lazyLoadImage}" id="discoverableCard-imag
bgset="https://c1.iggcdn.com/indiegogo-media-prod-c
load/c_fill,f_auto,h_273,w_273/gp5aurjxy2arn8nbflwu
style="background-image: url("https://c1.iggcdn.com
media-prod-cld/image/upload/c_fill,f_auto,h_273,w_2
y2arn8nbflwu.jpg");">…

<!-- ngIf: viewModel.isCampaign() -->
<div ng-if="viewModel.isCampaign()" class="disc
d-type discoverableCard-type--crowdfunding ng-sco
flex
<!-- end ngIf: viewModel.isCampaign() -->
<div class="discoverableCard-title ng-binding d
Card-lineClamp2" ng-class="::viewModel.titleCla
gogo-test="title" ng-bind="::viewModel.discover
e">SIRUI 75mm Anamorphic Lens King of Portrait<
<div class="discoverableCard-description ng-bin
erableCard-lineClamp3" ng-class="viewModel.descri
lass()" ng-bind="::viewModel.discoverable.tagline
<div class="discoverableCard-category ng-bindin
test="category" ng-click="clickCategory($event)
"::viewModel.discoverable.category">Camera Gear
<!-- ngIf: viewModel.showCrowdFundingProgress()
<div ng-if="viewModel.showCrowdFundingProgress(
iscoverableCard-crowdfundingProgress ng-scope">…<
<!-- end ngIf: viewModel.showCrowdFundingProgre
CRTL+F
… d a div.discoverableCard-body div.discoverableCard-title.ng-binding.discoverableCard-li
Styles
Computed
Layout
Event Listeners
DOM Breakpoints
Properties
Accessi
Filter
:hov .c
element.style {
}
@media (min-width: 480px)
.discoverableCard-title {
font-family: "benton-sans", "Helvetica", "sans-serif";
font-size: 16px;


$369,258 USD raised 73,852%
29 days left
$4,067 USD raised 813%
60 days left
$76,502 USD raised 3,825%
30 days left
30% OFF
div.discoverableCard-title.ng-binding.discoverableCard-lineClamp2 221.98 × 57
AirOgo Pilloon Ultralight - Most Versatile Jacket
Ready for any situation anywhere: daily commute, business travels, and your everyday activities
TRAVEL & OUTDOORS
$6,529 USD raised 131%
30 days left
TANGIBLE HOPE project
ING
Tangible Hope Project
Real People. Real Communities. Real Impact. A Documentary Series About Real Hope in America.
WEB SERIES & TV SHOWS
$2,450 USD raised 25%
30 days left
LIVALL
LIVALL EVO21 Smart Helmet
• 270° rear light • 350g ultra lightweight • SOS system
FUNDING
LIVALL EVO21 Smart Helmet: 360 Active...
The world's smartest lighting helmet with 270° rear light, patented fall detection and SOS alert
TRAVEL & OUTDOORS
$213,411 USD raised 2,134%
18 days left
FUNDED IN 15 MINUTES
PAGEPIERCE
Funded in 15 Minutes
ONEXPLAYER


media-prod-cld/image/upload/c_fill,f_auto,h_273,w_273/d3opvravwyfswbmvjlbh.jpg");">…

<!-- ngIf: viewModel.isCampaign() -->
…
flex
<!-- end ngIf: viewModel.isCampaign() -->
AirOgo Pilloon Ultralight – Most Versatile Jacket
…
Travel & Outdoors

: viewModel.showCrowdFundingProgress() -->
f="viewModel.showCrowdFundingProgress()" class="discoverableCard-crowdfundingProgress ng-scope">…
<!-- end ngIf: viewModel.showCrowdFundingProgress() -->
<!-- ngIf: viewModel.isStandardCampaign() -->
.discoverableCard-title
… ng-isolate-scope div.discoverableCard a div.discoverableCard-body div.discoverableCard-title.ng-
.discoverableCard-title 5 of 13 Cancel
Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility
Filter :hov .cls
element.style {
}
@media (min-width: 480px)
.discoverableCard-title {
font-family: "benton-sans", "Helvetica", "sans-serif";
font-size: 16px;
<style>
Console What's New
Highlights from the Chrome 90 update

Techniques - XPATH

XPath is a syntax for defining parts of an XML document.
XPath uses path expressions to navigate in XML documents.

```
<bookstore>

  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>

  <book category="children">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>

...
```

/bookstore/book/title

Everyday Italian
Harry Potter

https://www.w3schools.com/xml/xpath_examples.asp

Expression	Description
nodename	Selects all nodes with the name "*nodename*"
/	Selects from the root node
//	Selects nodes in the document from the current node that match the selection no matter where they are
.	Selects the current node
..	Selects the parent of the current node
@	Selects attributes

Wildcard	Description
*	Matches any element node
@*	Matches any attribute node
node()	Matches any node of any kind

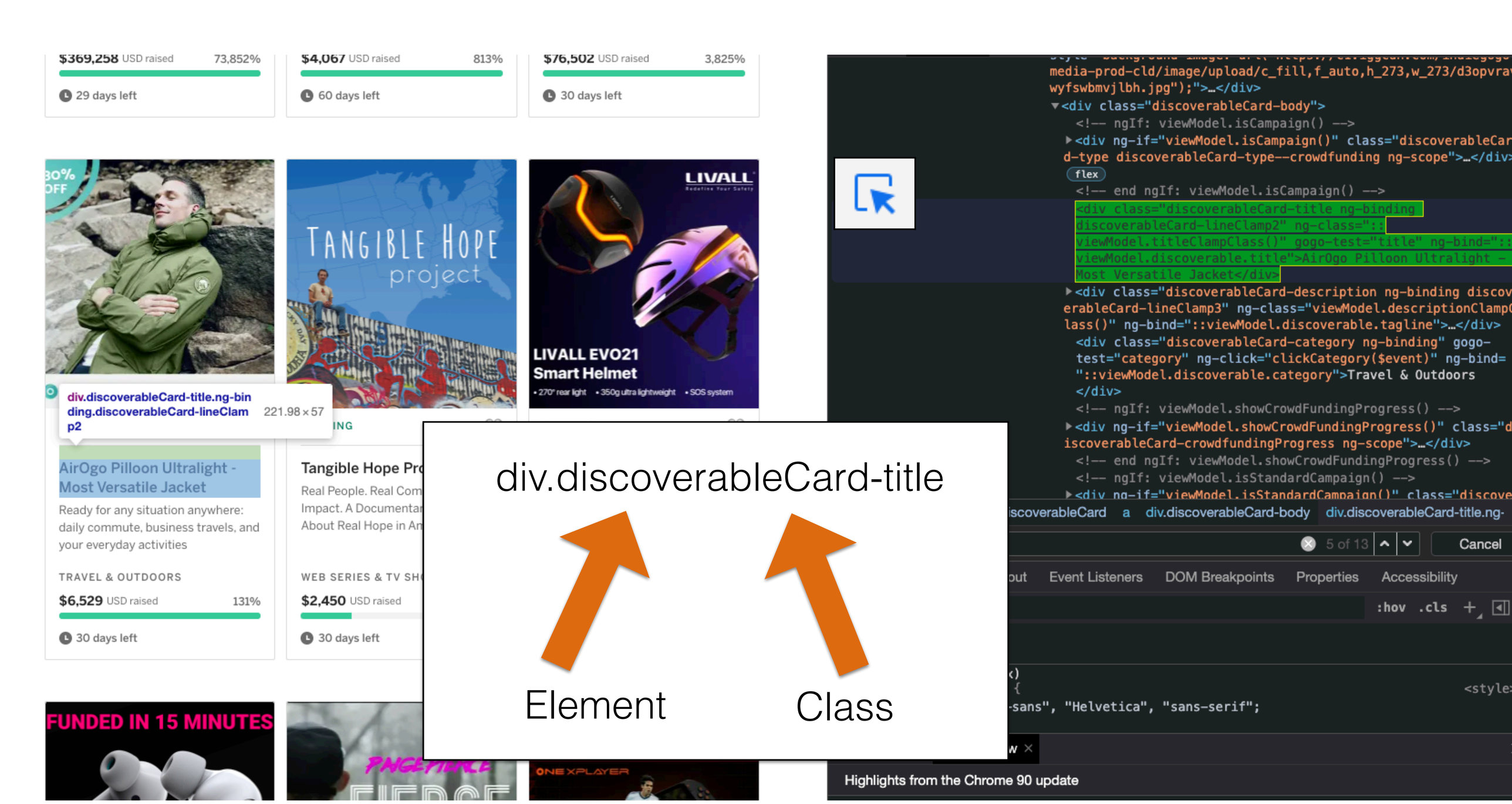
Techniques - CSS Selector

In CSS, selectors declare which part of the markup a style applies to by **matching tags and attributes** in the markup itself.

Selector overview

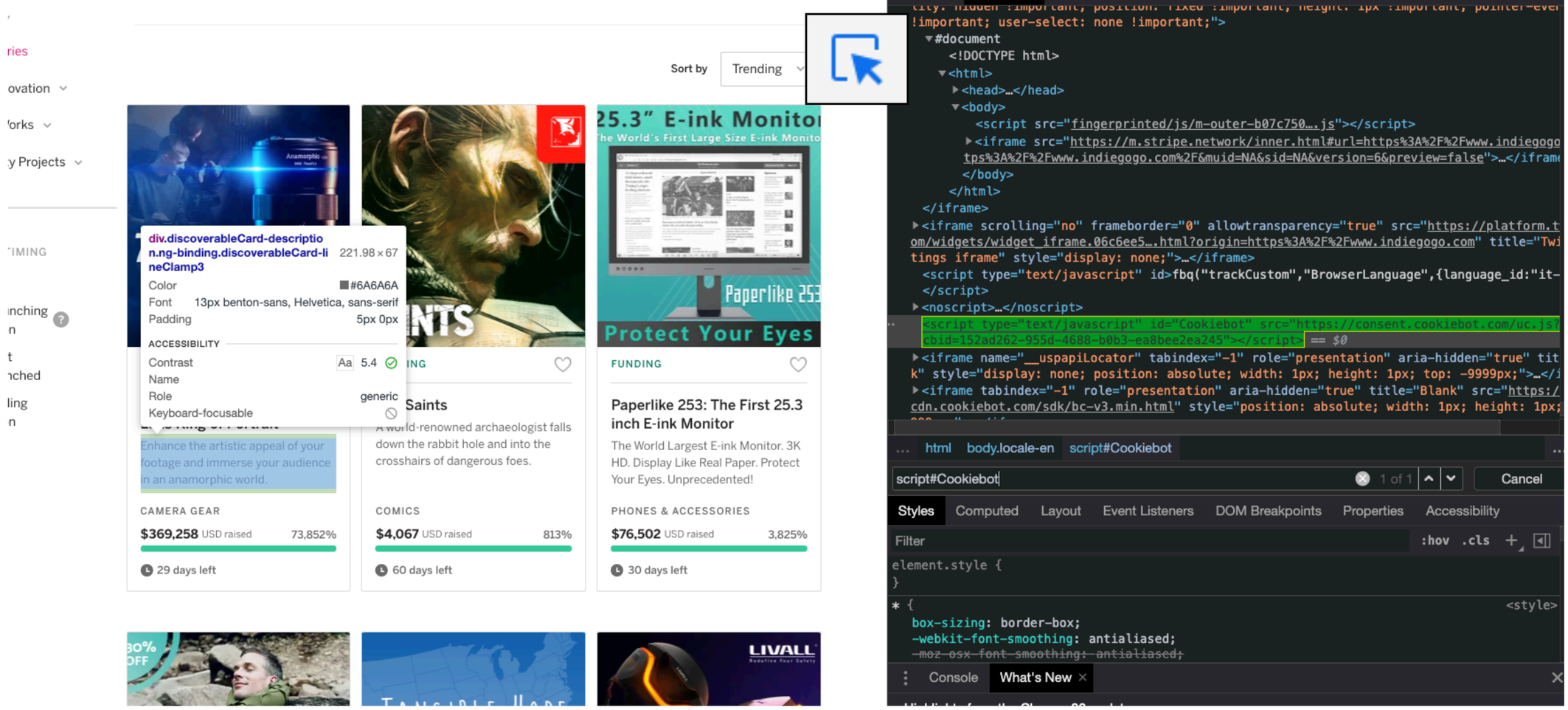
- `tagname`: find elements by tag, e.g. `a`
- `ns|tag`: find elements by tag in a namespace, e.g. `fb|name` finds `<fb:name>` elements
- `#id`: find elements by ID, e.g. `#logo`
- `.class`: find elements by class name, e.g. `.masthead`
- `[attribute]`: elements with attribute, e.g. `[href]`
- `[^attr]`: elements with an attribute name prefix, e.g. `[^data-]` finds elements with HTML5 dataset attributes
- `[attr=value]`: elements with attribute value, e.g. `[width=500]` (also quotable, like `[data-name='launch sequence']`)
- `[attr^=value]`, `[attr$=value]`, `[attr*=value]`: elements with attributes that start with, end with, or contain the value, e.g. `[href*=/path/]`
- `[attr~=regex]`: elements with attribute values that match the regular expression; e.g. `img[src~=(?i)\.(png|jpe?g)]`
- `*`: all elements, e.g. `*`

https://www.w3schools.com/cssref/css_selectors.asp


$369,258 USD raised 73,852%
29 days left
$4,067 USD raised 813%
60 days left
$76,502 USD raised 3,825%
30 days left
TANGIBLE HOPE project
LIVALL
LIVALL EVO21
Smart Helmet
• 270° rear light • 350g ultra lightweight • SOS system
div.discoverableCard-title.ng-binding.discoverableCard-lineClamp2
221.98 × 57
AirOgo Pilloon Ultralight - Most Versatile Jacket
Ready for any situation anywhere: daily commute, business travels, and your everyday activities
TRAVEL & OUTDOORS
$6,529 USD raised 131%
30 days left
Tangible Hope Pro
Real People. Real Com
Impact. A Documentar
About Real Hope in An
WEB SERIES & TV SH
$2,450 USD raised
30 days left
FUNDED IN 15 MINUTES
PAGEFIERCE
ONEXPLAYER
media-prod-cld/image/upload/c_fill,f_auto,h_273,w_273/d3opvrawyfswbmvjlbh.jpg");">…

<!-- ngIf: viewModel.isCampaign() -->
…
flex
<!-- end ngIf: viewModel.isCampaign() -->
AirOgo Pilloon Ultralight - Most Versatile Jacket
…
Travel & Outdoors

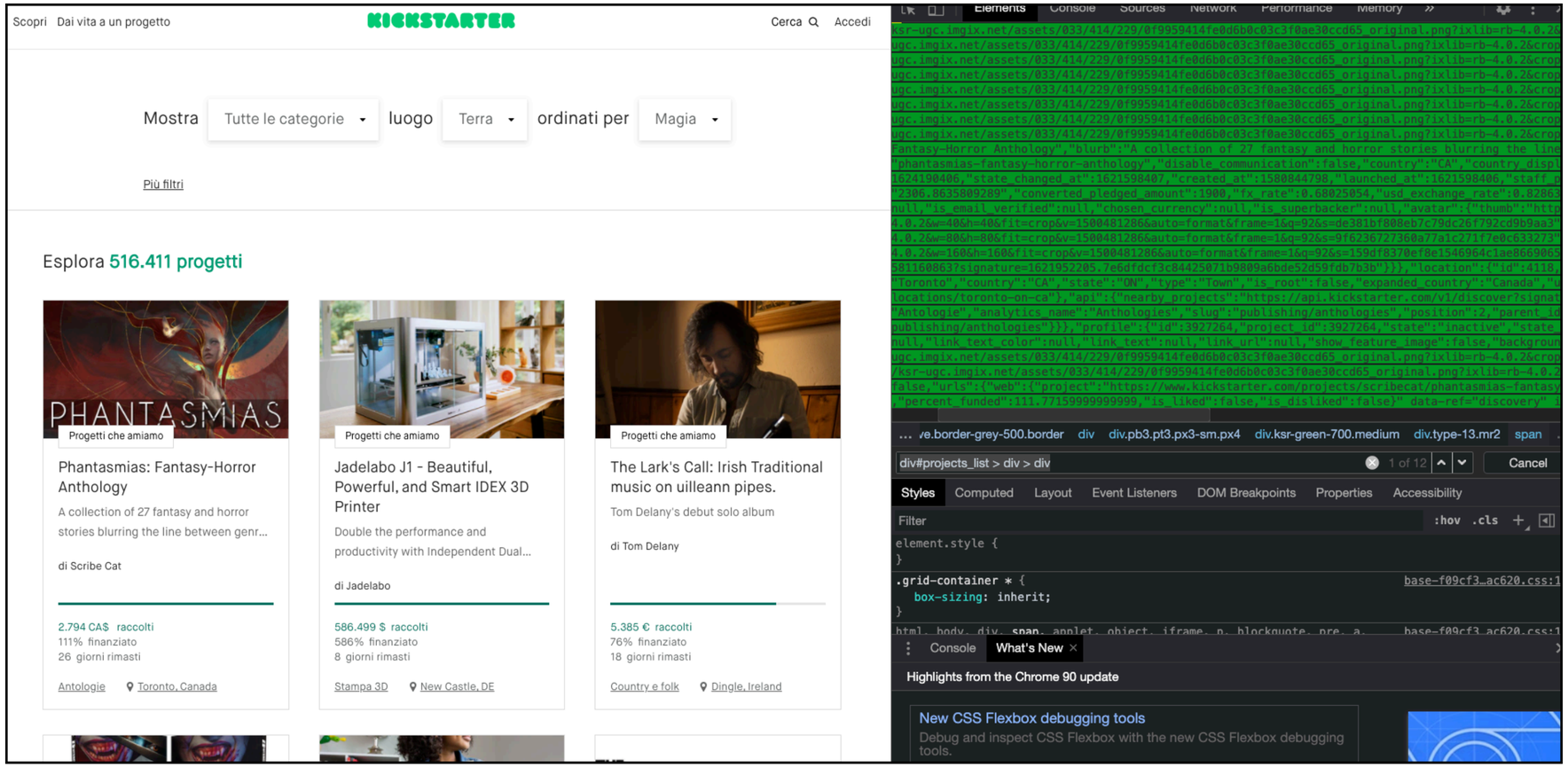
<!-- ngIf: viewModel.showCrowdFundingProgress() -->
…
<!-- end ngIf: viewModel.showCrowdFundingProgress() -->
<!-- ngIf: viewModel.isStandardCampaign() -->
iscoverableCard a div.discoverableCard-body div.discoverableCard-title.ng-
5 of 13
Cancel
out Event Listeners DOM Breakpoints Properties Accessibility
:hov .cls
-sans", "Helvetica", "sans-serif";
Highlights from the Chrome 90 update
div.discoverableCard-title
Element
Class



script#Cookiebot

Element

ID



discoverable-card > div > a

By parent relation

Ready for the challenge?

How can I select **all titles in the list**?

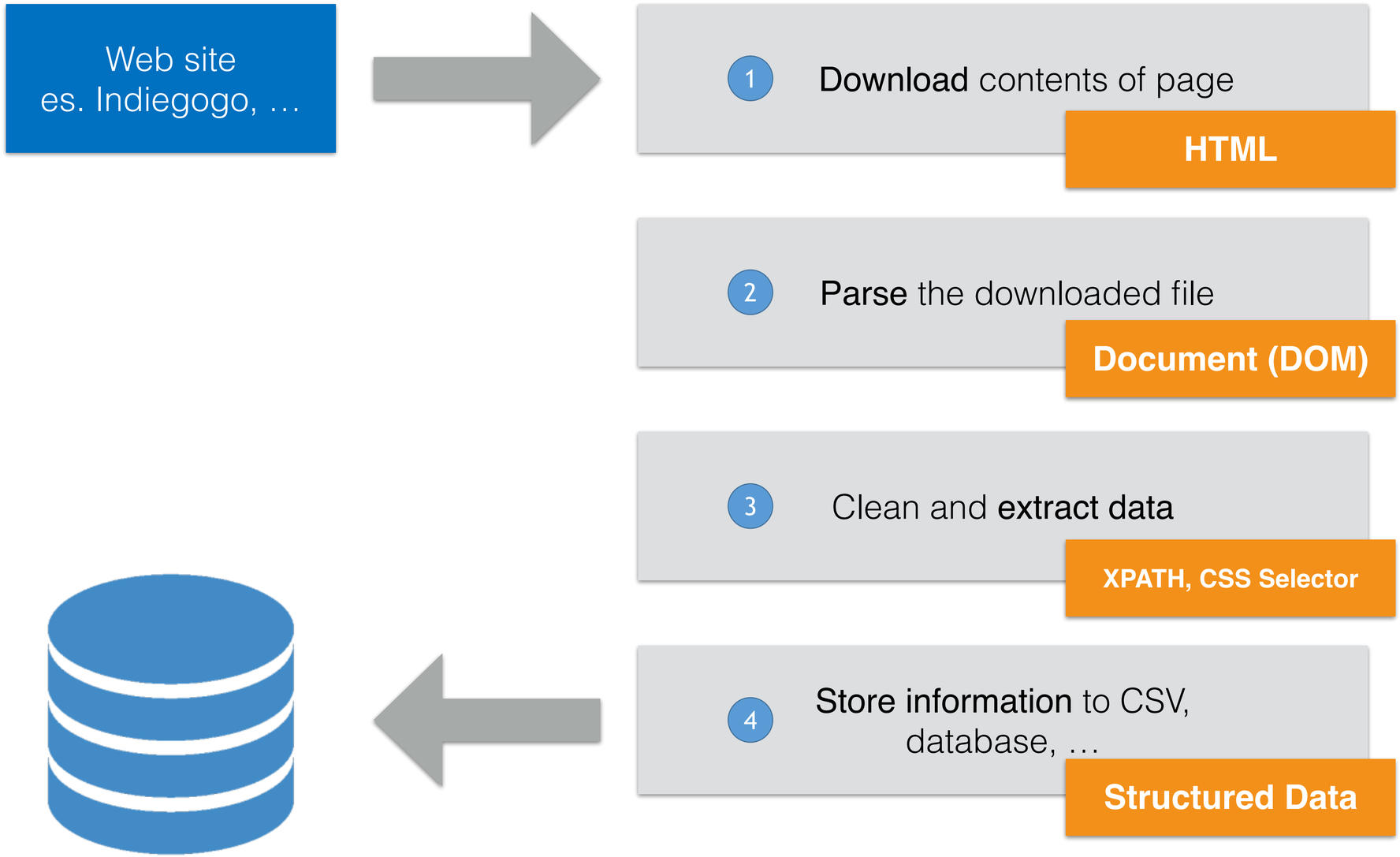
A. List > Title

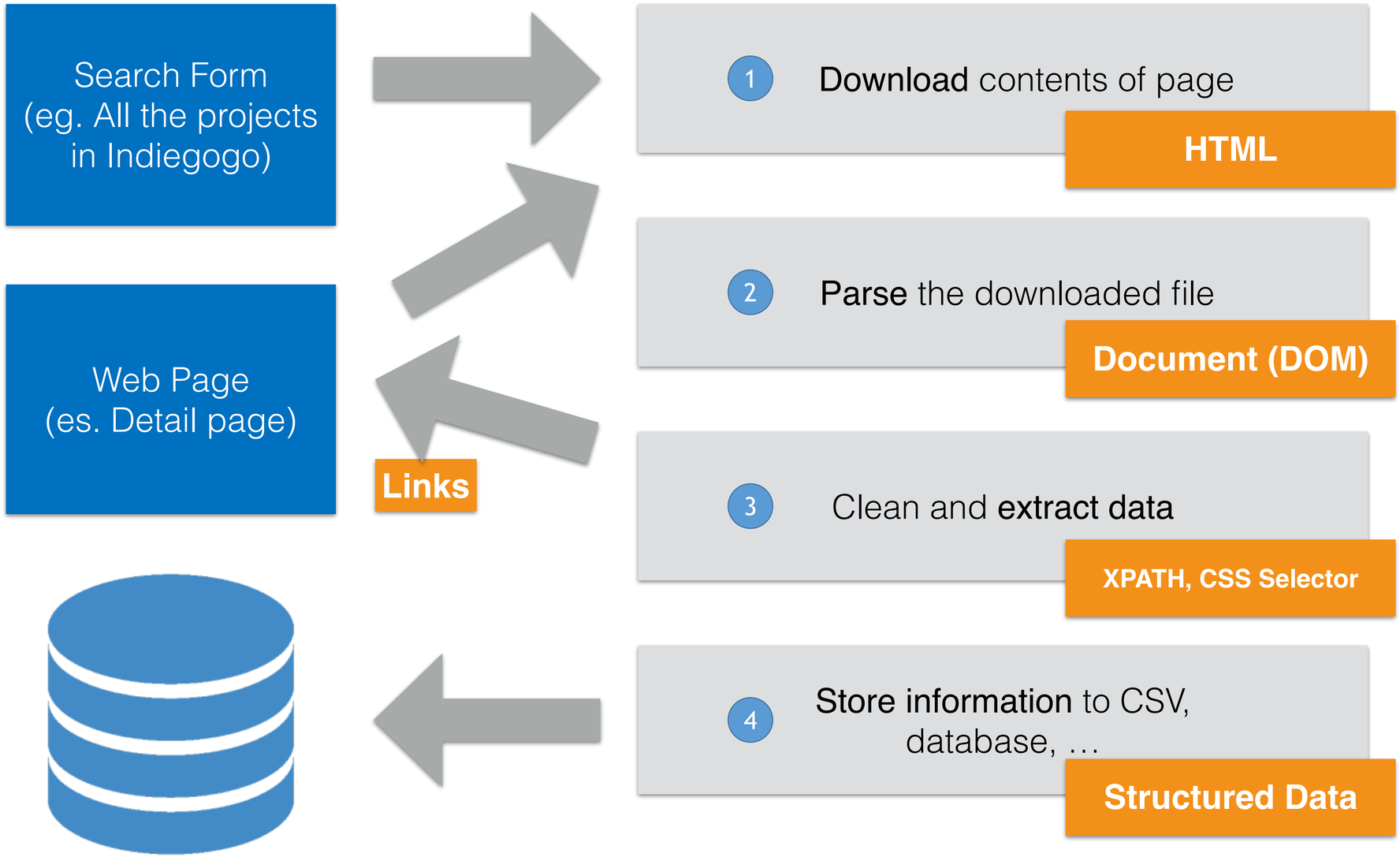
B. table > td

C. discoverable-card

D. discoverable-card > .discoverableCard-title

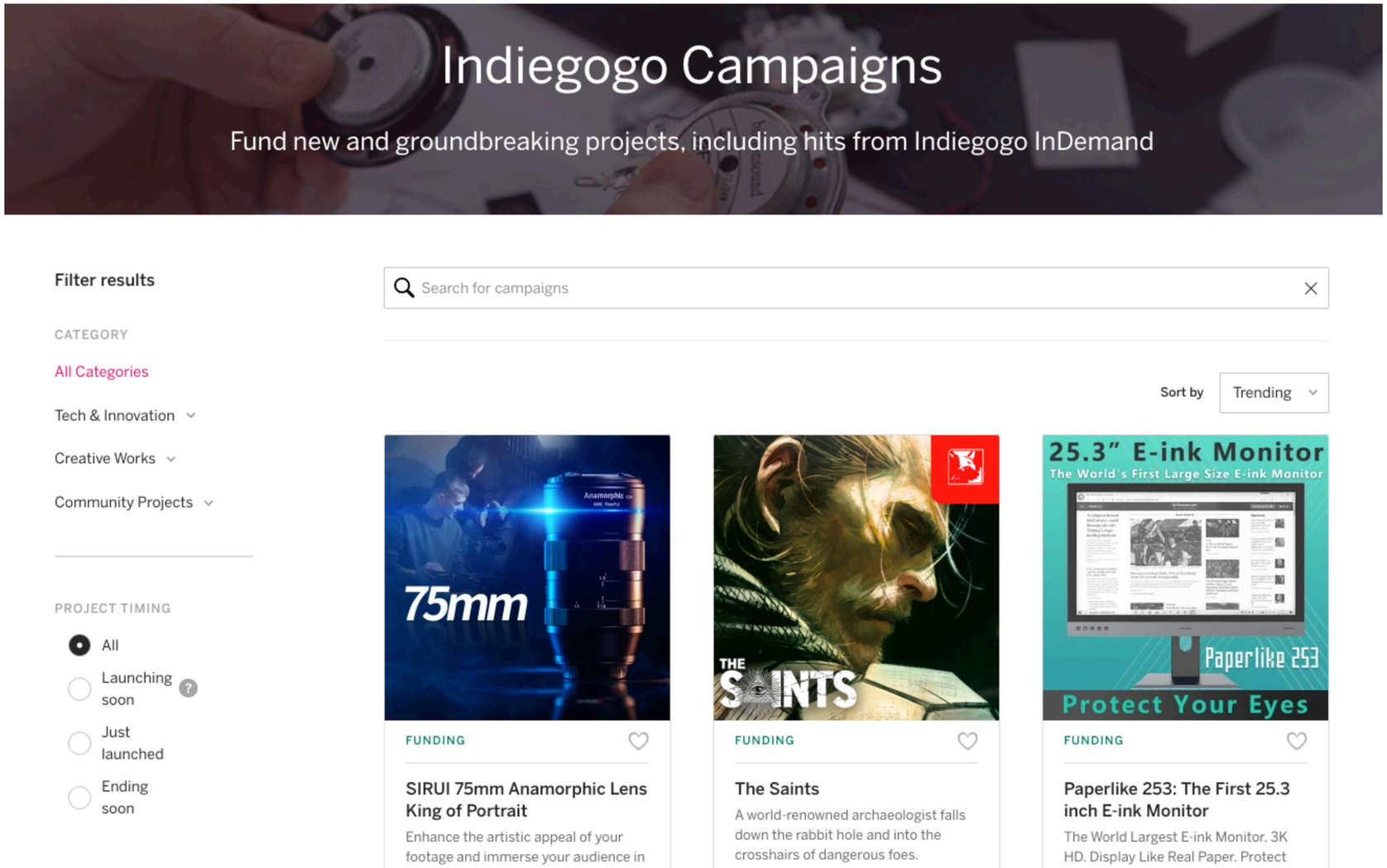
E. discoverable-card .discoverableCard-title


Web site
es. Indiegogo, …
1
Download contents of page
HTML
2
Parse the downloaded file
Document (DOM)
3
Clean and extract data
XPATH, CSS Selector
4
Store information to CSV,
database, …
Structured Data


Search Form
(eg. All the projects
in Indiegogo)
Web Page
(es. Detail page)
Links
1
Download contents of page
HTML
2
Parse the downloaded file
Document (DOM)
3
Clean and extract data
XPATH, CSS Selector
4
Store information to CSV,
database, …
Structured Data

Inspecting a website

https://www.indiegogo.com/explore/all?project_type=campaign&project_timing=all&sort=trending



Inspecting a website

FUNDING

AirOgo Pilloon Ultralight - Most Versatile Jacket

Ready for any situation anywhere: daily commute, business travels, and your everyday activities

TRAVEL & OUTDOORS

$6,529 USD raised 131%

30 days left

FUNDING

Tangible Hope Project

Real People. Real Communities. Real Impact. A Documentary Series About Real Hope in America.

WEB SERIES & TV SHOWS

$2,450 USD raised 25%

30 days left

FUNDING

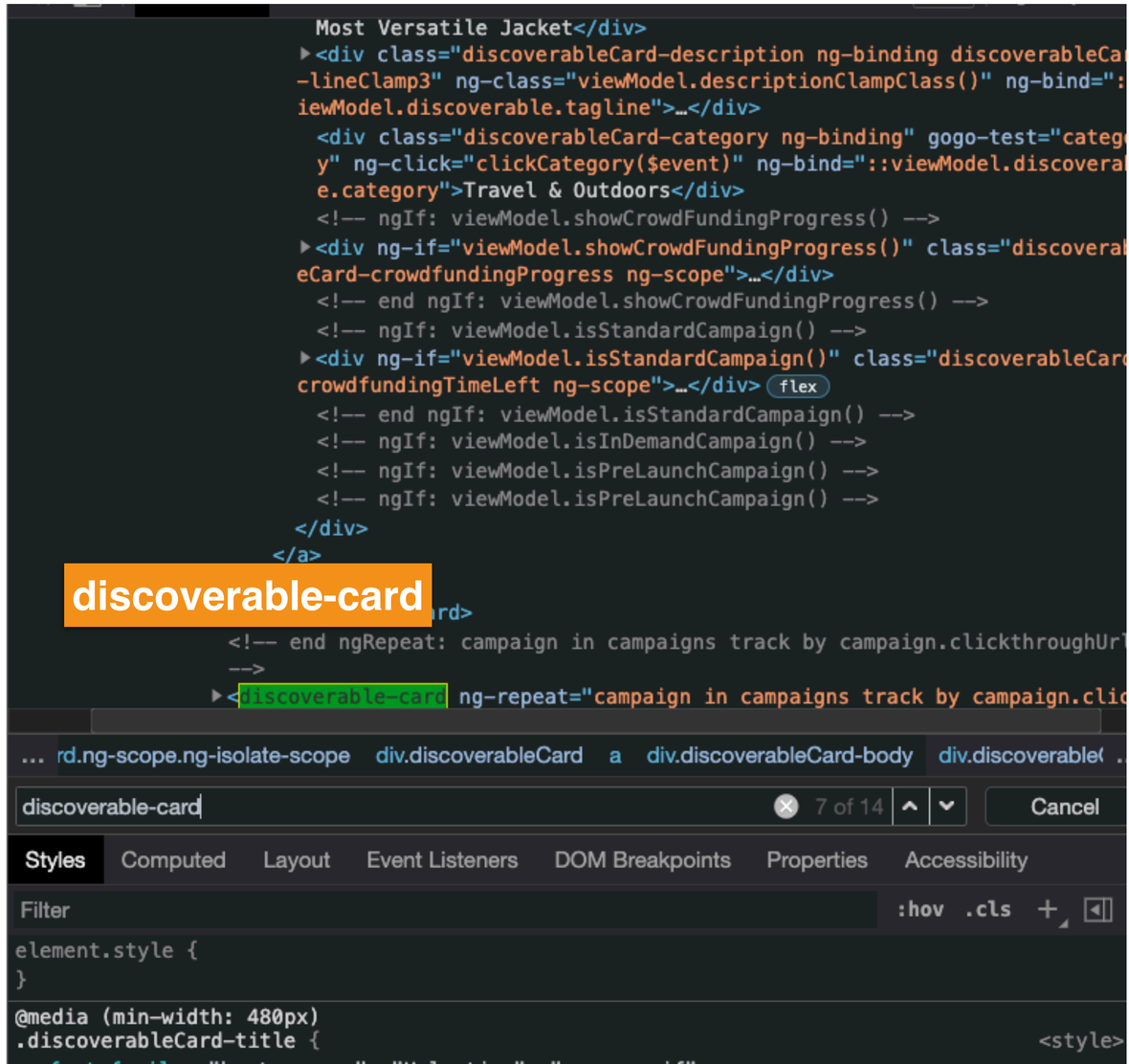
LIVALL EVO21 Smart Helmet: 360 Active...

The world's smartest lighting helmet with 270° rear light, patented fall detection and SOS alert

TRAVEL & OUTDOORS

$213,411 USD raised 2,134%

18 days left



Inspecting a website

FUNDING

AirOgo Pilloon Ultralight - Most Versatile Jacket

Ready for any situation anywhere: daily commute, business travels, and your everyday activities

TRAVEL & OUTDOORS

$6,529 USD raised 131%

30 days left

FUNDING

Tangible Hope Project

Real People. Real Communities. Real Impact. A Documentary Series About Real Hope in America.

WEB SERIES & TV SHOWS

$2,450 USD raised 25%

30 days left

FUNDING

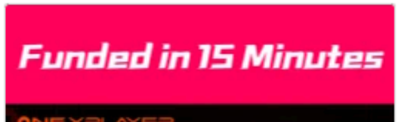
LIVALL EVO21 Smart Helmet: 360 Active...

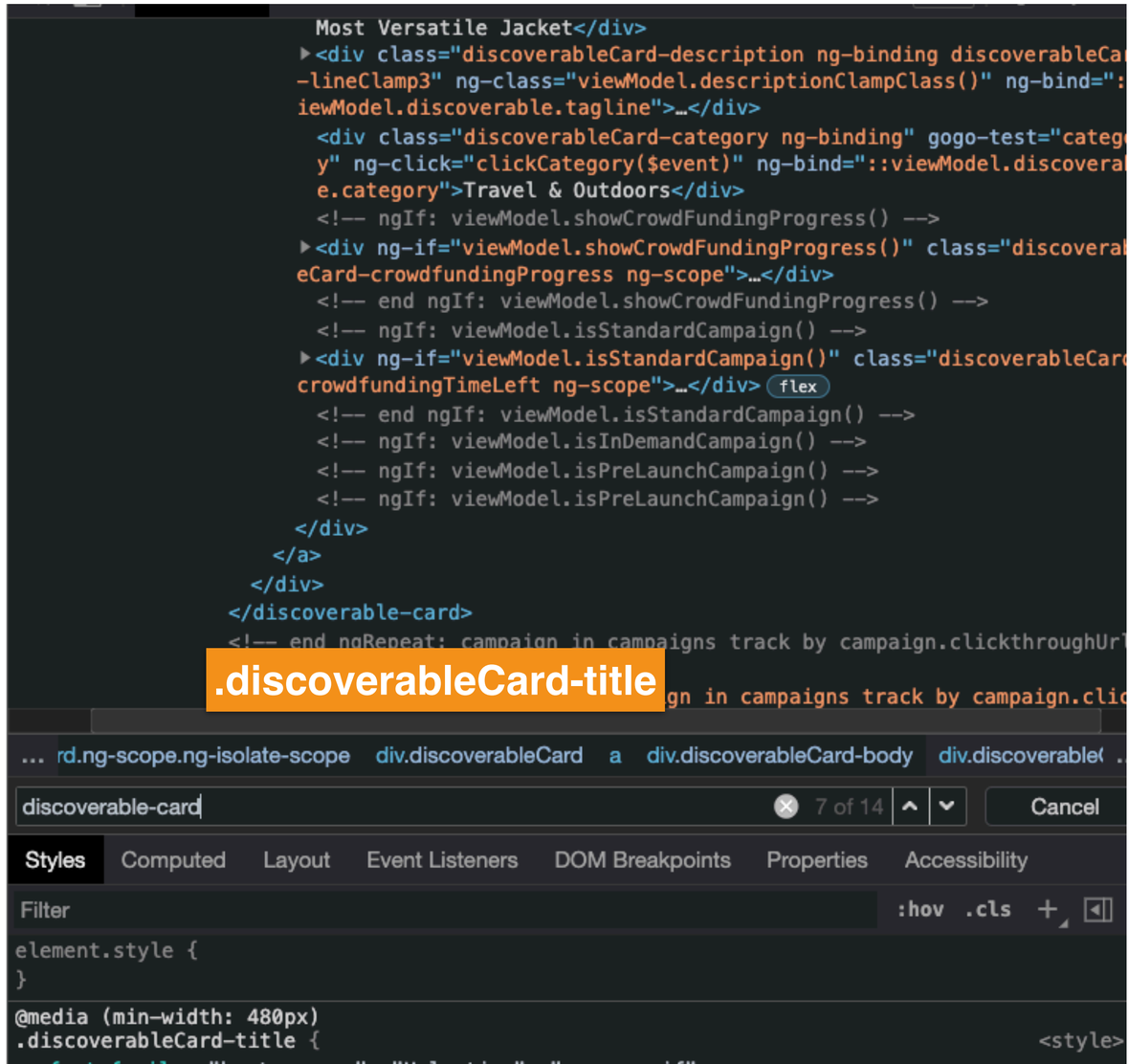
The world's smartest lighting helmet with 270° rear light, patented fall detection and SOS alert

TRAVEL & OUTDOORS

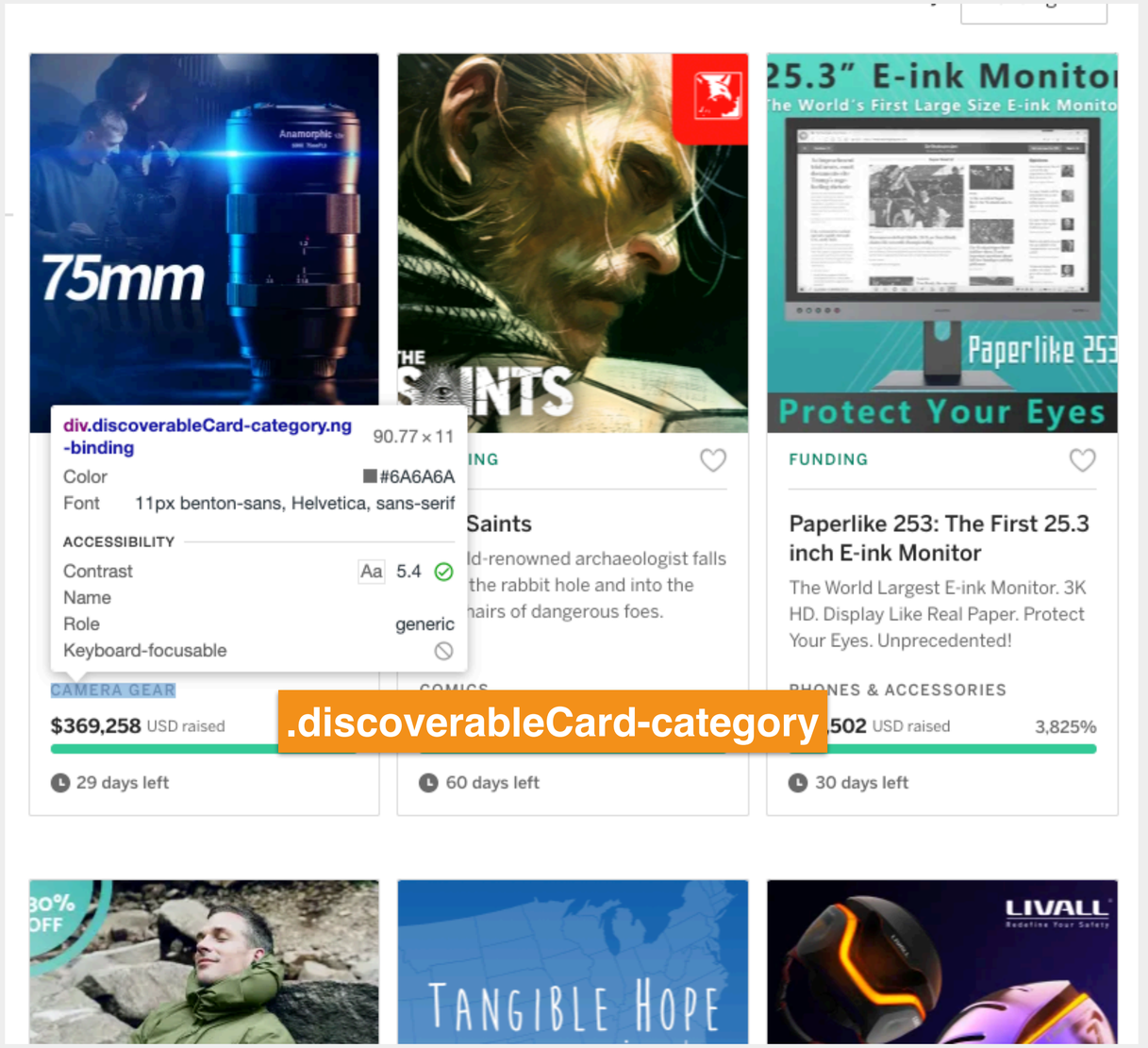
$213,411 USD raised 2,134%

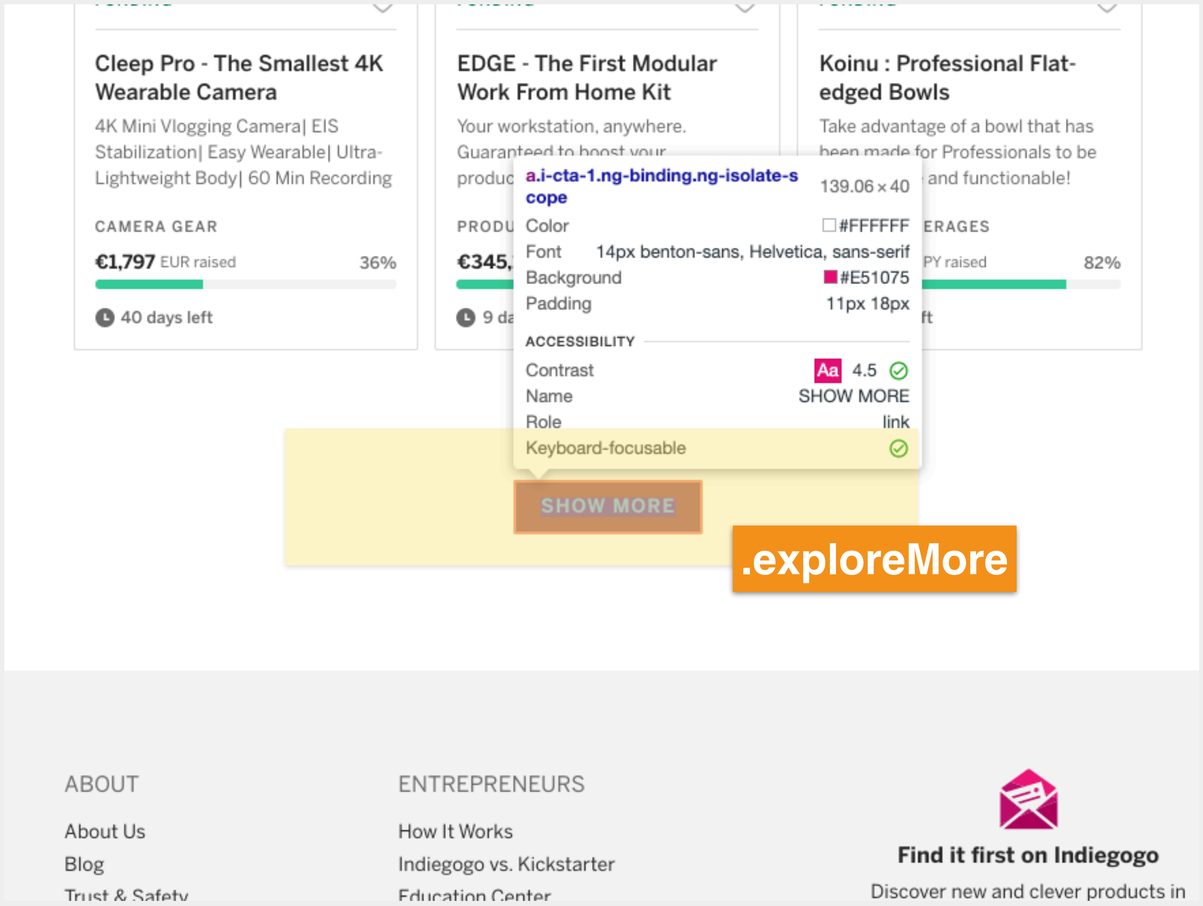
18 days left





Inspecting a website





Inspecting a website

Recap

Link
discoverable-card a [href]

Category
.discoverableCard-category

Title
.discoverableCard-title

Description
.discoverableCard-description

BeautifulSoup

A toolkit for dissecting a document and extracting what you need.

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

The examples in this documentation should work the same way in Python 2.7 and Python 3.2.

You might be looking for the documentation for Beautiful Soup 3. If so, you should know that Beautiful Soup 3 is no longer being developed, and that Beautiful Soup 4 is recommended for all new projects. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see Porting code to BS4.

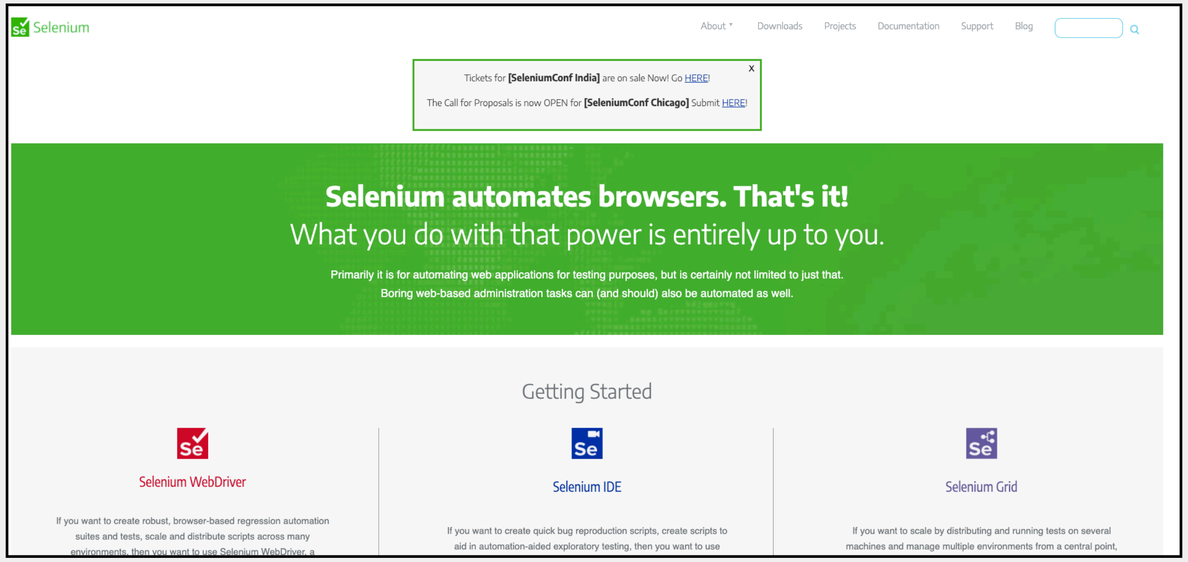
```
pip install --user  beautifulsoup4
pip install --user  requests
pip install --user  tqdm
```

Selenium

Selenium is a portable framework for testing web applications.

Scraping with Selenium

https://www.selenium.dev/



Are you ready?

Colab

https://bit.ly/3Bvuq43

JSOUP

jsoup: Java HTML Parser

`jsoup` is a Java library for working with real-world HTML. It provides a very convenient API for extracting and manipulating data, using the best of DOM, CSS, and jquery-like methods.

`jsoup` implements the WHATWG HTML5 specification, and parses HTML to the same DOM as modern browsers do.

- scrape and parse HTML from a URL, file, or string
- find and extract data, using DOM traversal or CSS selectors
- manipulate the HTML elements, attributes, and text
- clean user-submitted content against a safe white-list, to prevent XSS attacks
- output tidy HTML

jsoup is designed to deal with all varieties of HTML found in the wild; from pristine and validating, to invalid tag-soup; jsoup will create a sensible parse tree.

pip install scrapy
scrapy startproject zalando
scrapy shell <url>
scarpa crawl zalando -o file.csv

Scrapy

https://docs.scrapy.org/en/latest/intro/tutorial.html

Ethics

Supreme Court revives LinkedIn case to protect user data from web scrapers

Zack Whittaker @zackwhittaker / 8:54 PM GMT+2 • June 14, 2021

Comment

Facebook: Stolen Data Scraped from Platform in 2019

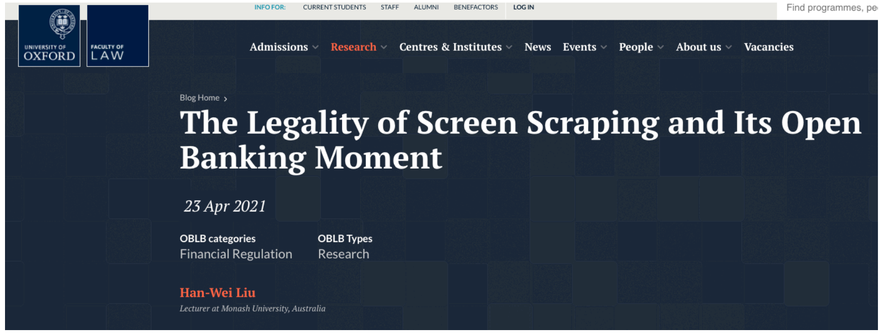
Author:
Elizabeth Montalbano
April 7, 2021 / 9:00 am
3 minute read
Write a comment
Share this article:

The flaw that caused the leak of personal data of more than 533 million users over the weekend no longer exists; however, the social media giant still faces an investigation by EU regulators.

Scraping Away at the CFAA– The Supreme Court's Interpretation of "Exceeds Authorized Access" Limits the Scope of the Statute's Application to Data Scrapers

June 17, 2021

In a long awaited decision that has a significant application for data scraping, the Supreme Court issued on June 3, 2021 its decision in *Van Buren v. United States*, significantly limiting the scope of the Computer Fraud and Abuse Act by holding that users who access information that they are entitled to obtain but use that information for improper purposes do not violate the statute. The majority opinion adopted a narrow interpretation of the statute that will make it more difficult to pursue both civil and criminal actions based on alleged misappropriation of data.



MIT Technology Review

Featured Topics Newsletters Events Podcasts

Opinion

Web scraping is a tool, not a crime

Web scraping is one of the most powerful tools journalists have to hold companies and governments accountable.

by **Lam Thuy Vo** December 8, 2020

Ethics

- Some websites do not like scraping. Why?
  - Increase load on web servers
  - Private websites
  - Dynamic websites
  - Data ownership
- It is important to "be recognised
  - No fraudulent behaviour (e.g. we insert a sleep(x) in our scripts)
  - Notify
  - Ask for the data (are we sure it is not already available in a structured format?)
- Always check API policies
  - Limits
  - Rate