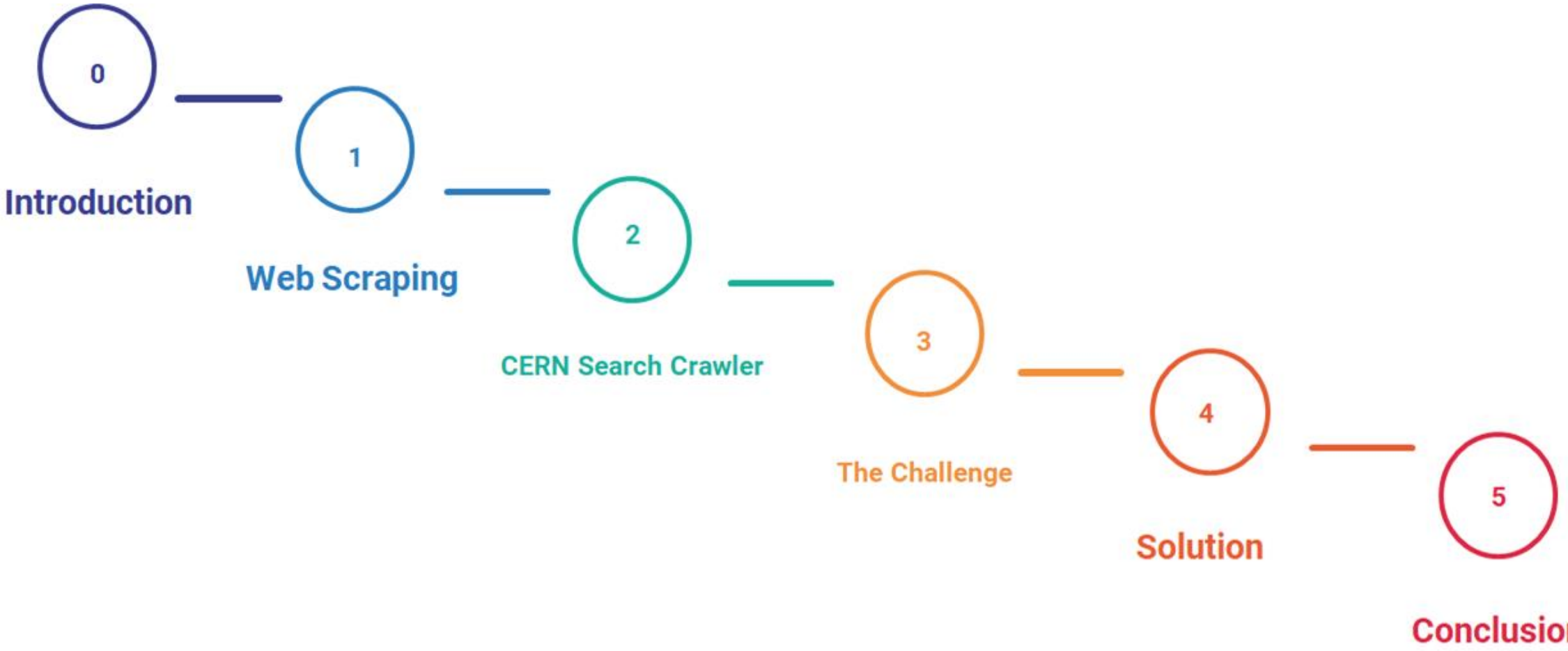
CERN openlab

CERN Search web crawler

Supporting JavaScript rendered websites

Speaker: **khansa Amrouni**
Supervisors: **Pablo Panero**, **Andreas Wagner**
Department: **IT-CDA-WF**

Plan


0
Introduction
1
Web Scraping
2
CERN Search Crawler
3
The Challenge
4
Solution
5
Conclusion

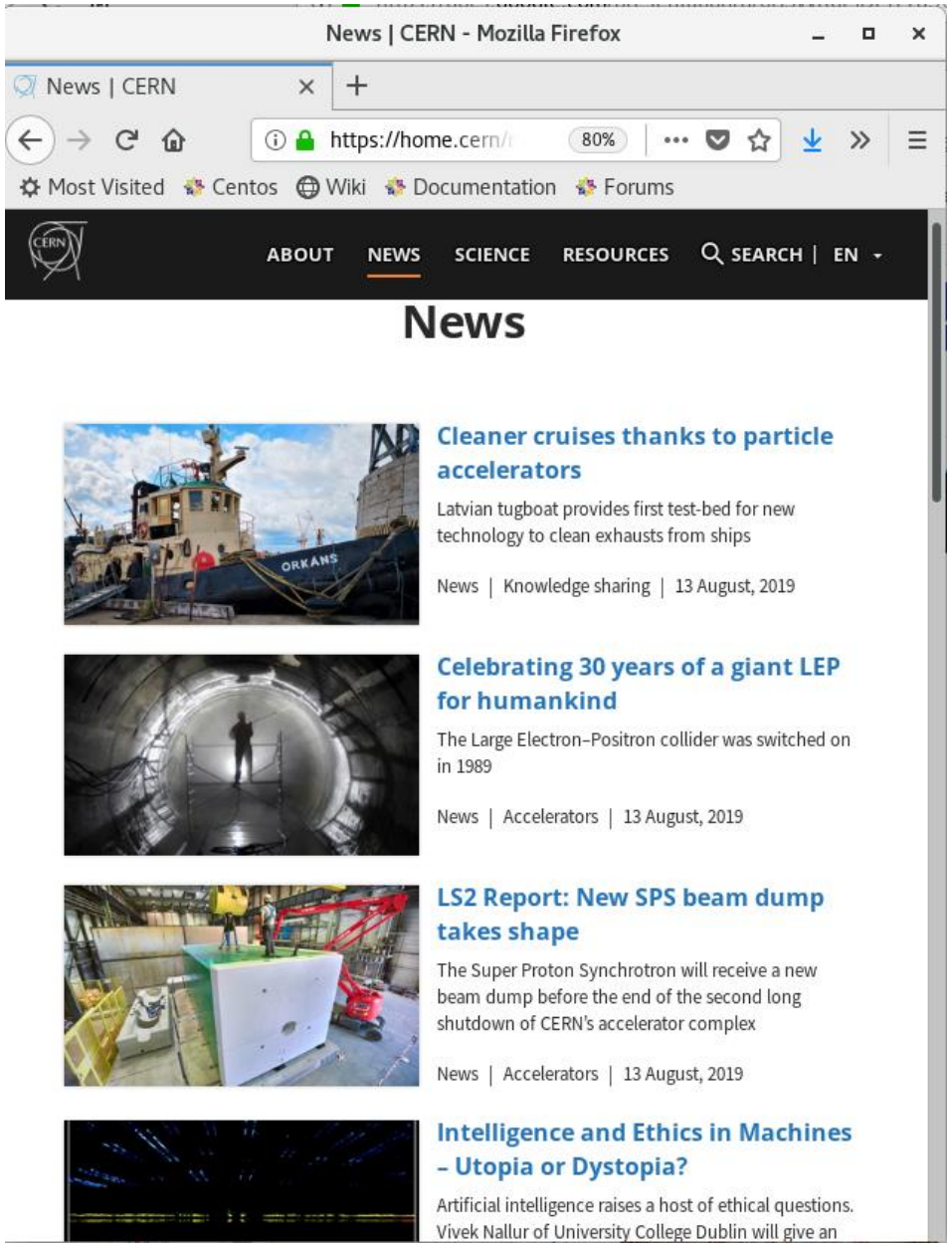
Introduction

Web Scraping

Web Scraping

Web Scraping is a technique for **extracting** the **content of websites** and present it in different format.

Web Scraping: example



	Title	Content	Tags
2	Cleaner cruises thanks to particle accelerators	Latvian tugboat provides first test-bed for new technology to c	News, Knowledge sharing
3	Celebrating 30 years of a giant LEP for humankind	The Large Electron–Positron collider was switched on in 198	News, Accelerators
4	LS2 Report: New SPS beam dump takes shape	The Super Proton Synchrotron will receive a new beam dump	News, Accelerators
5	Intelligence and Ethics in Machines – Utopia or Dys	Artificial intelligence raises a host of ethical questions. Vivek N	Announcement, Computing
6	Closure of Route Fermi in the direction of Entrance	N/A	Announcement, At CERN
7	Defibrillators and emergency equipment in the LHC	In the event of a medical emergency in the tunnel during LS2,	Announcement, At CERN
8	Ready, Steady, Goal: Table Football Tournament cc	32 teams. 127 matches. 1 winning team. What a week it was!	News, At CERN
9	The 2018 CERN Annual Report is available	You can read it online or get a paper copy at the Library	Announcement, At CERN
10	CERN theorist shares Special Breakthrough Prize i	Sergio Ferrara, Daniel Z. Freedman and Peter van Nieuwenh	News, Knowledge sharing
11	Sixty years of the CERN Courier	The magazine has published over 600 issues and now reache	News, Knowledge sharing
12	..	..	..
13			
14			

CSV

0 Introduction

1 Web Scraping

2 CERN Search Crawler

3 The Challenge

4 Solution

5 Conclusion

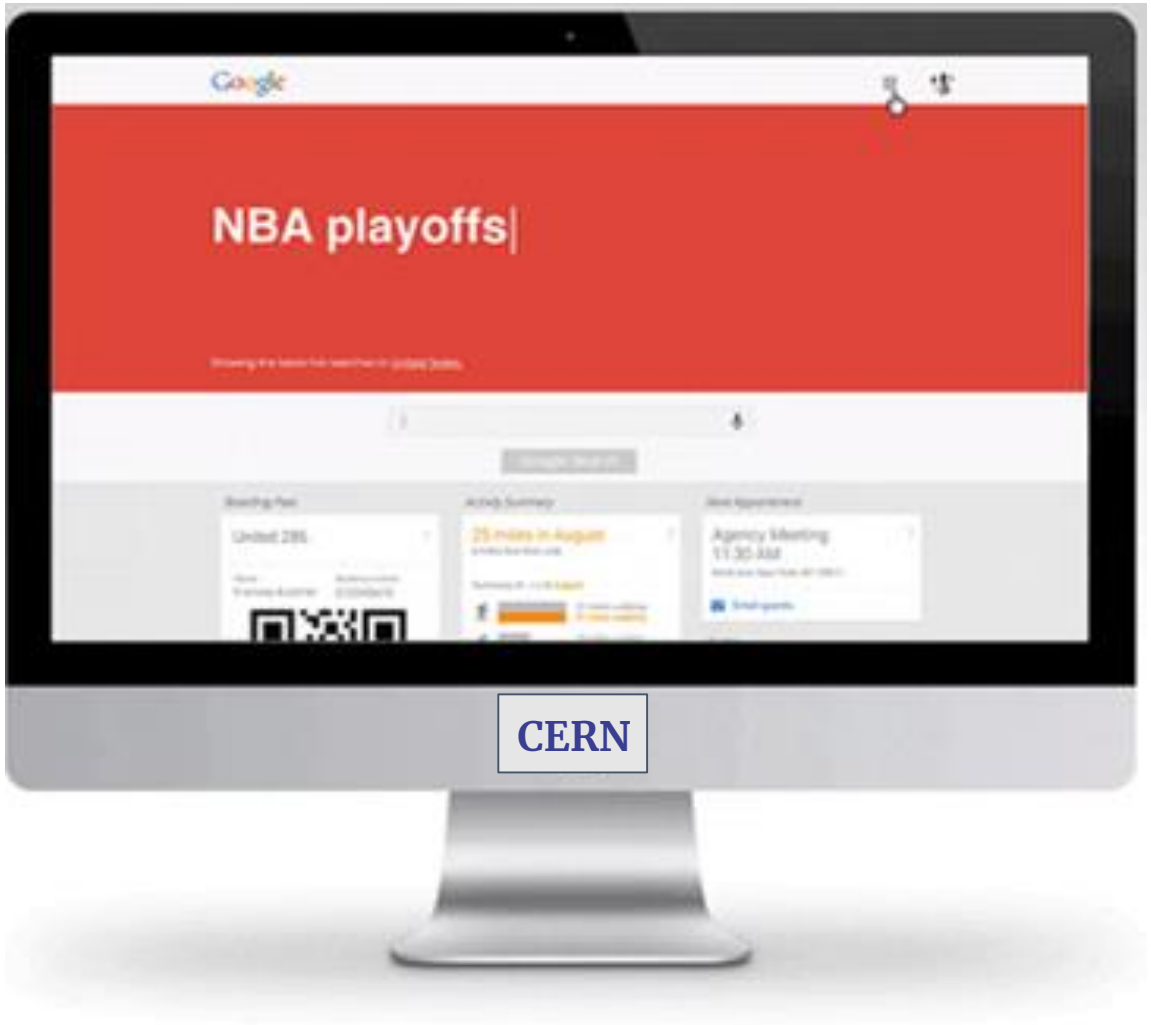
CERN Search Crawler

- The CERN Search Crawler is a focused crawler developed using **scrapy** for the cern.ch websphere.

- The goal of this crawler is **to feed** the **CERN Search Engine** with the content corresponding to all project websites hosted at CERN.

CERNsearch

The Challenge




```
<!DOCTYPE html>
<html lang="en">
<head>...</head>
<body>
      <noscript>You need to enable JavaScript to run this app.
      </noscript>
        <div id="root"></div>
        <script type="text/javascript"
                src="/static/js/main.6d8ab65a.js"> </script>

</body>
</html>
```

How we resolved it?

Splash

Lightweight, scriptable browser as a service with an HTTP API.

OPENSHIFT

open source

docker

But wait..

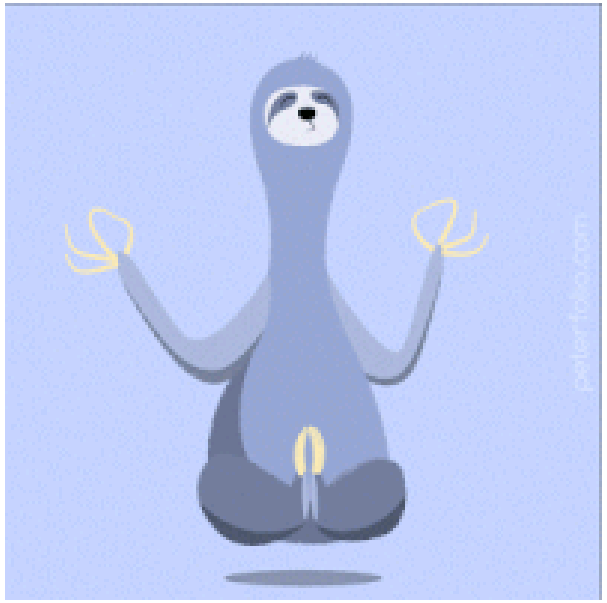
is it really efficient?

Let's TEST it..

	Only scrapy	Scrapy + Splash
Run Time (1 website)	~[4...5] s	~[11..26] s

it may slow down our scraping process ..

No worries: we have a solution



JS exist ?

no

yes

normal request

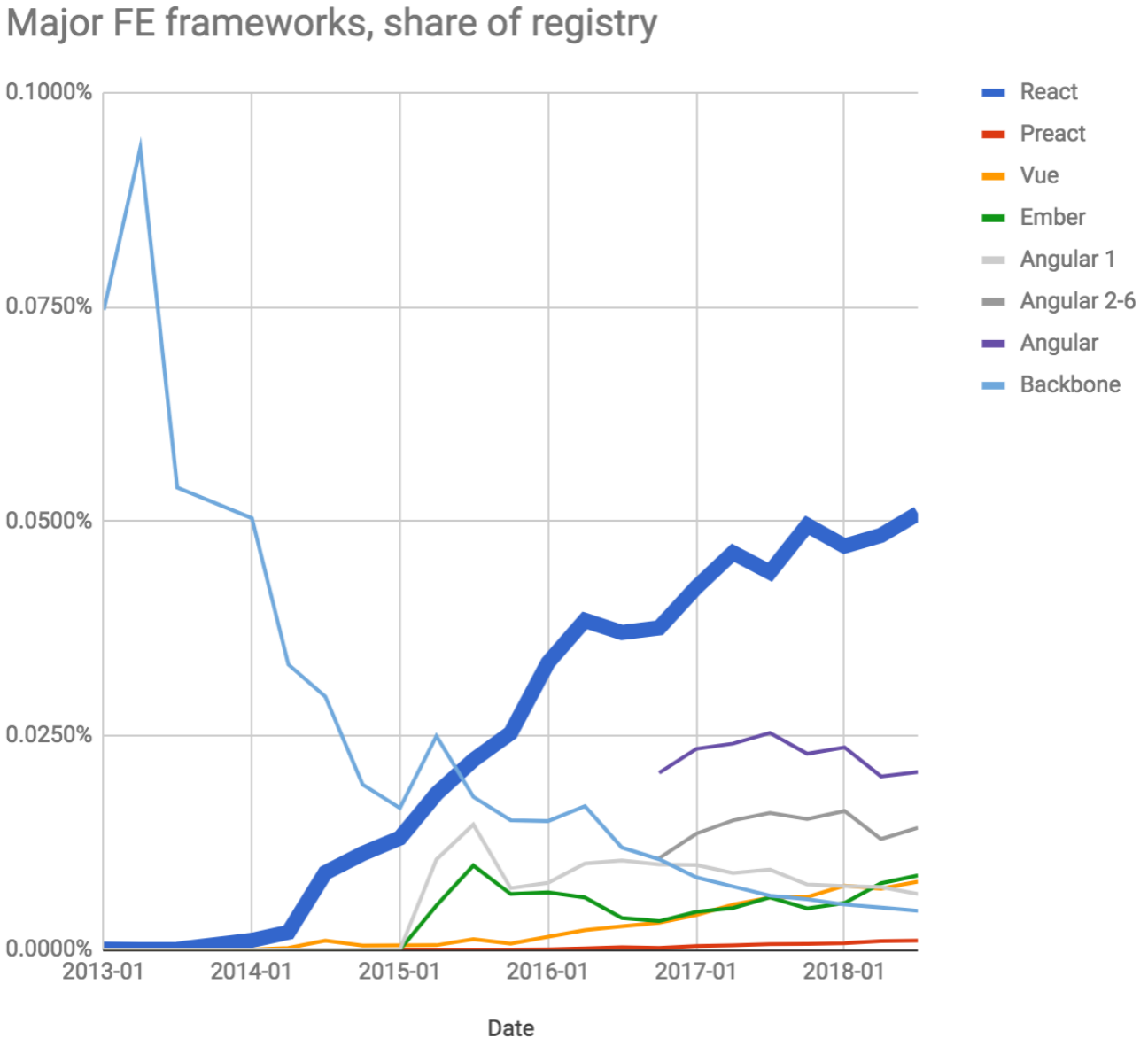
splash request

Conclusion

```
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
                var layer = L.marker(
                    [response[i].latitude, response[i].longitude]
                    // ,{icon: myIcon}
                );
                layer.addTo(group);

                layer.bindPopup(
                    "<p>" + "Species: " + response[i].species + "<
                    "<p>" + "Description: " + response[i].descript
                    "<p>" + "Seen at: " + response[i].latitude + '
                    "<p>" + "On: " + response[i].sighted_at + "</
            }

            $('select').change(function() {
                species = this.value;
            });
        });
    }
$.ajax({
            url: queryURL,
            method: "GET"
        }).done(function(response) {
            for (var i = 0; i < response.length; i++) {
                var layer = L.marker(
                    [response[i].latitude, response[i].longitude]
                    // ,{icon: myIcon}
                );
                layer.addTo(group);
```



This year in JavaScript: 2018 in review and npm's predictions for 2019, npm, Inc, Dec 6, 2018

References

- **Gitlab Cern Search Crawler project**, accessed **14,08,2019**
  link: **https://gitlab.cern.ch/webservices/cern-search/web-crawler**

- **CERN News websites**, accessed **14,08,2019**
  link: **https://home.cern/news**

- **GIFs website**, accessed **13,08,2019**, link: **https://giphy.com/**

- **Scrapy Documentation website**, accessed **11, 08, 2019**
  link: **https://scrapy.org/**

- Medium, **this-year-in-javascript-2018-in-review-and-npms-predictions-for-2019**, accessed **15, 08, 2019**,
  link:**https://medium.com/npm-inc/this-year-in-javascript-2018-in-review-and-npms-predictions-for-2019-3a3d7e5298ef**

Wants to know more about the project ?

Contact me!

khansa.amrouni@cern.ch

My website link: