

Fundamentos Ciência de Dados

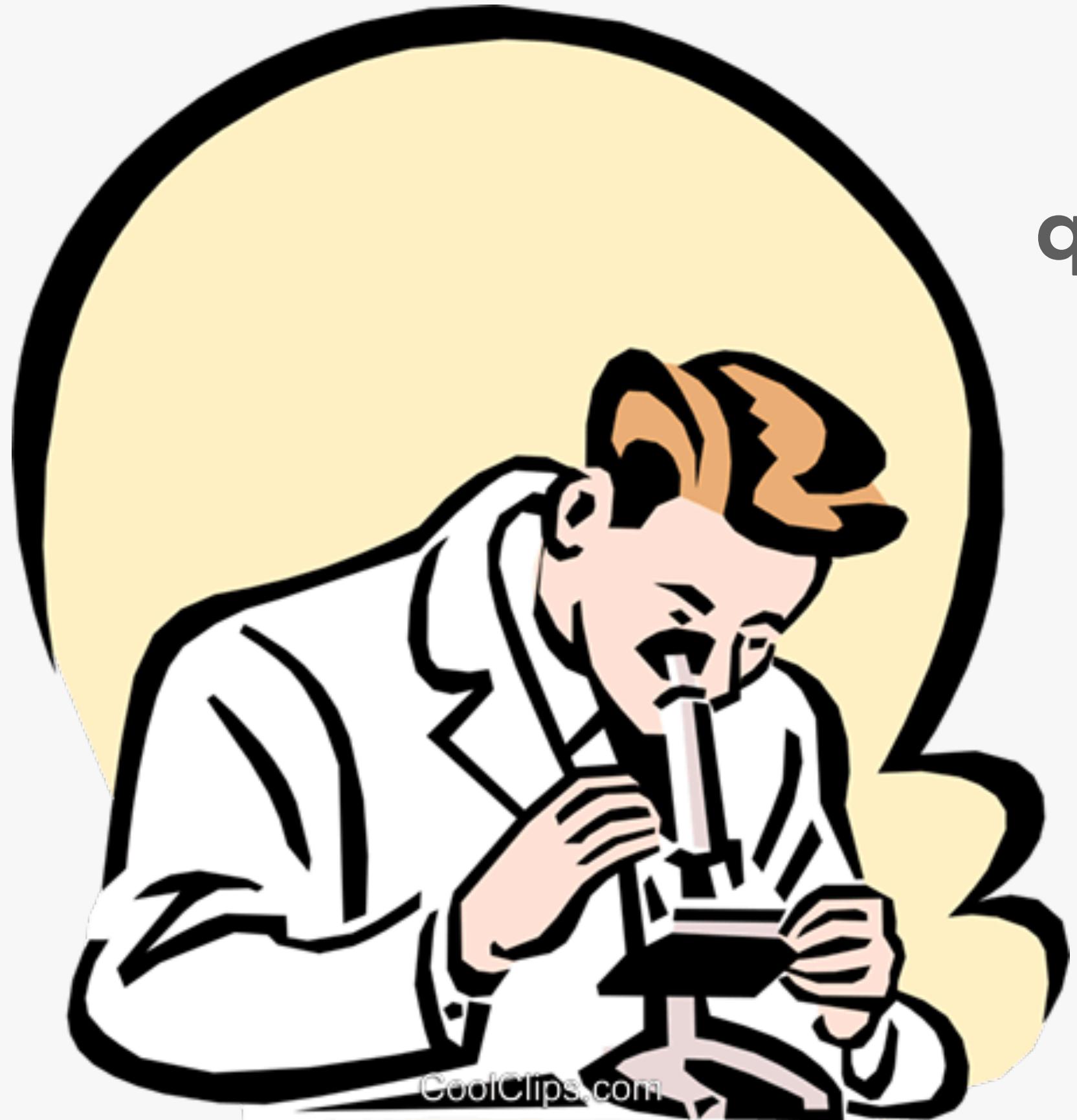
Scientific Recommender: Análise Inicial Sobre Artigos Publicados em Conferências

Alunos: Ingrid Pacheco, Eduardo Prata e Renan Parreira
Professores: Sérgio Serra e Jorge Zavatela

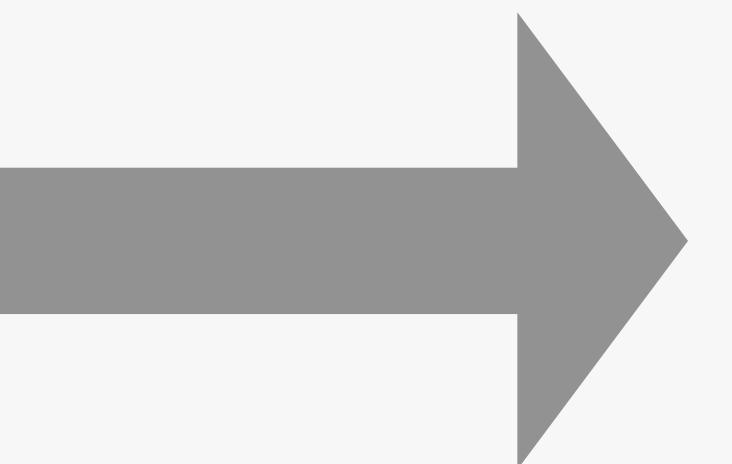
PROBLEMA INICIAL



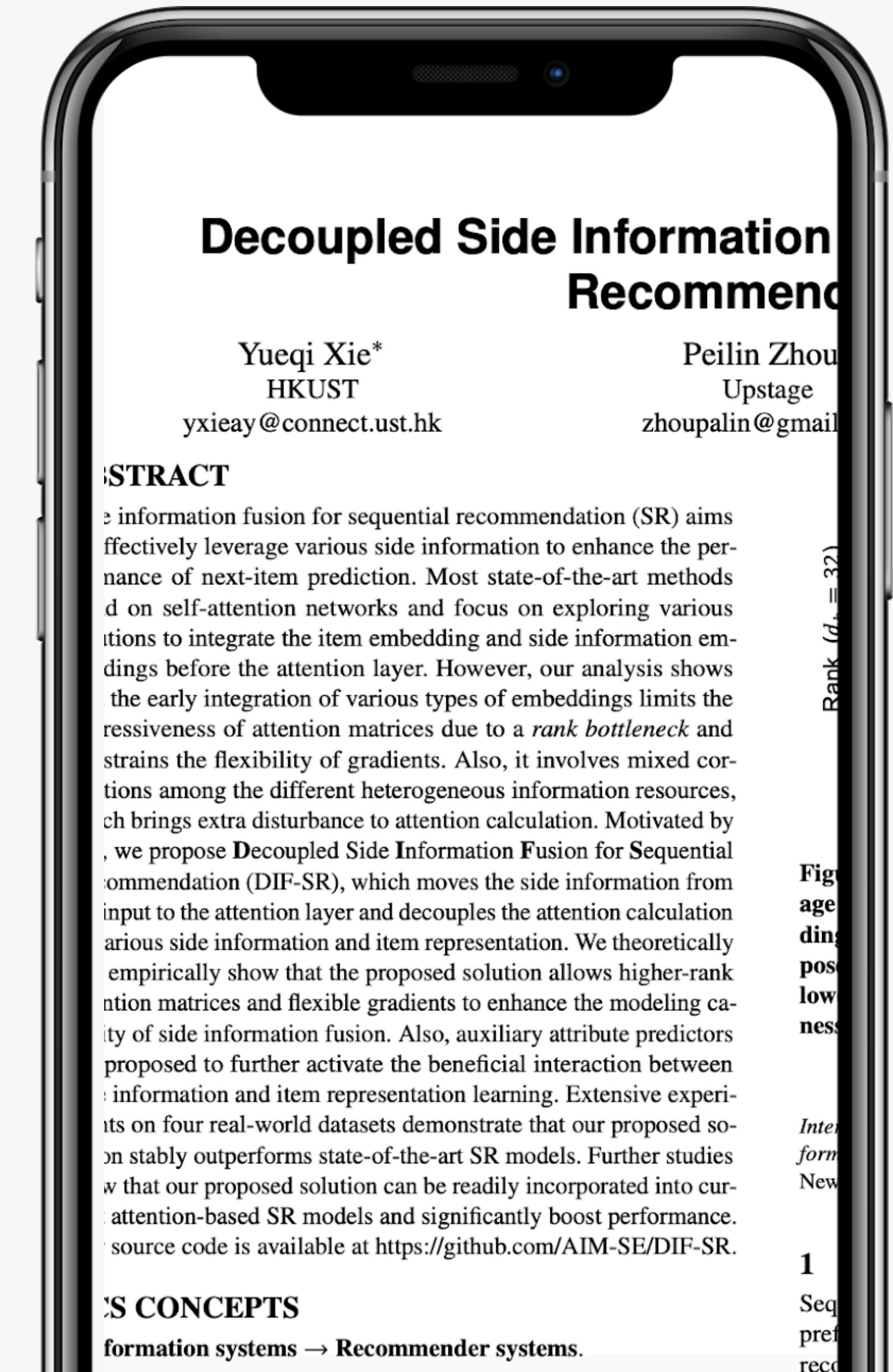
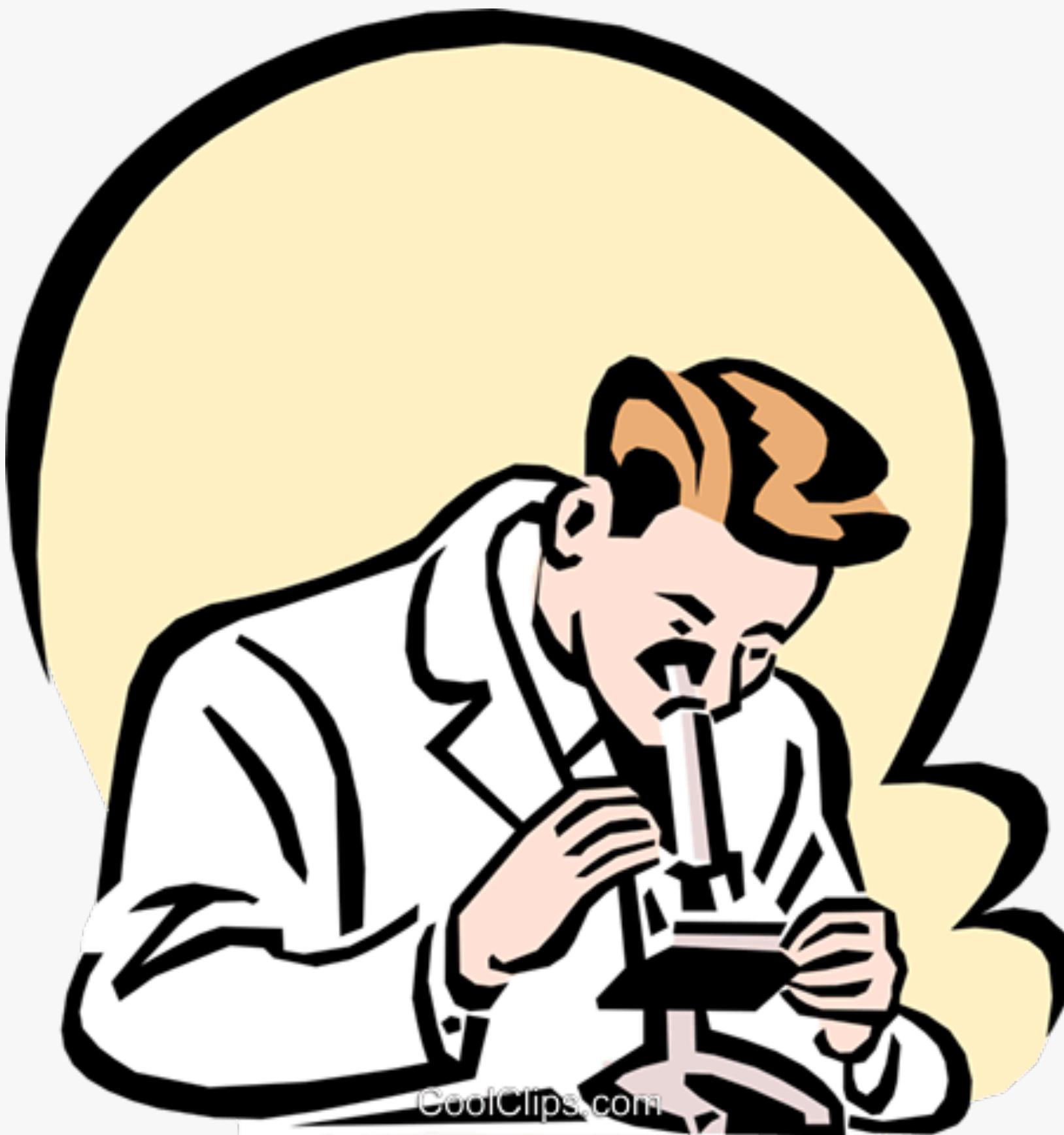
Problema Inicial



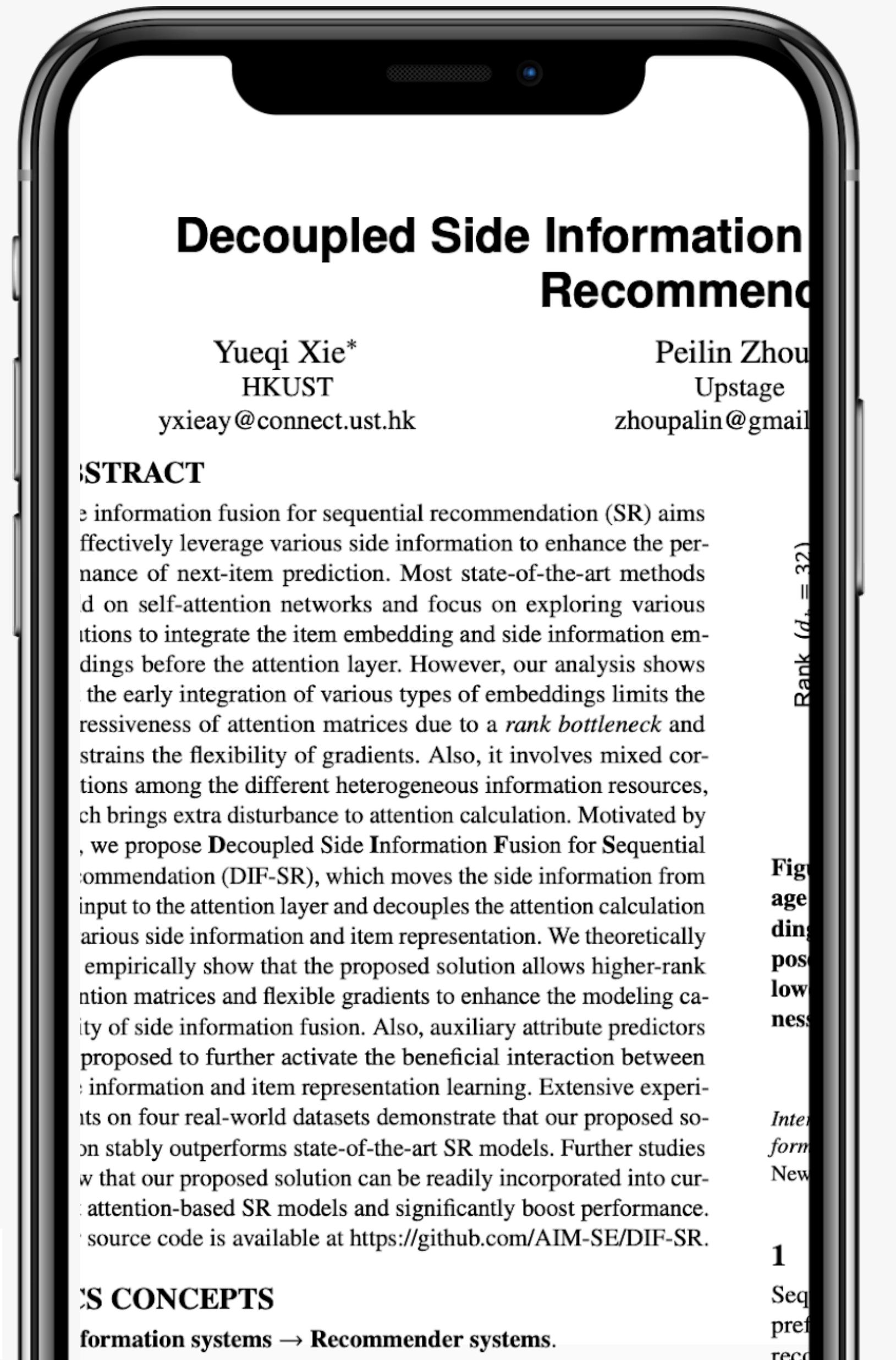
Maior
quantidade de
pessoas



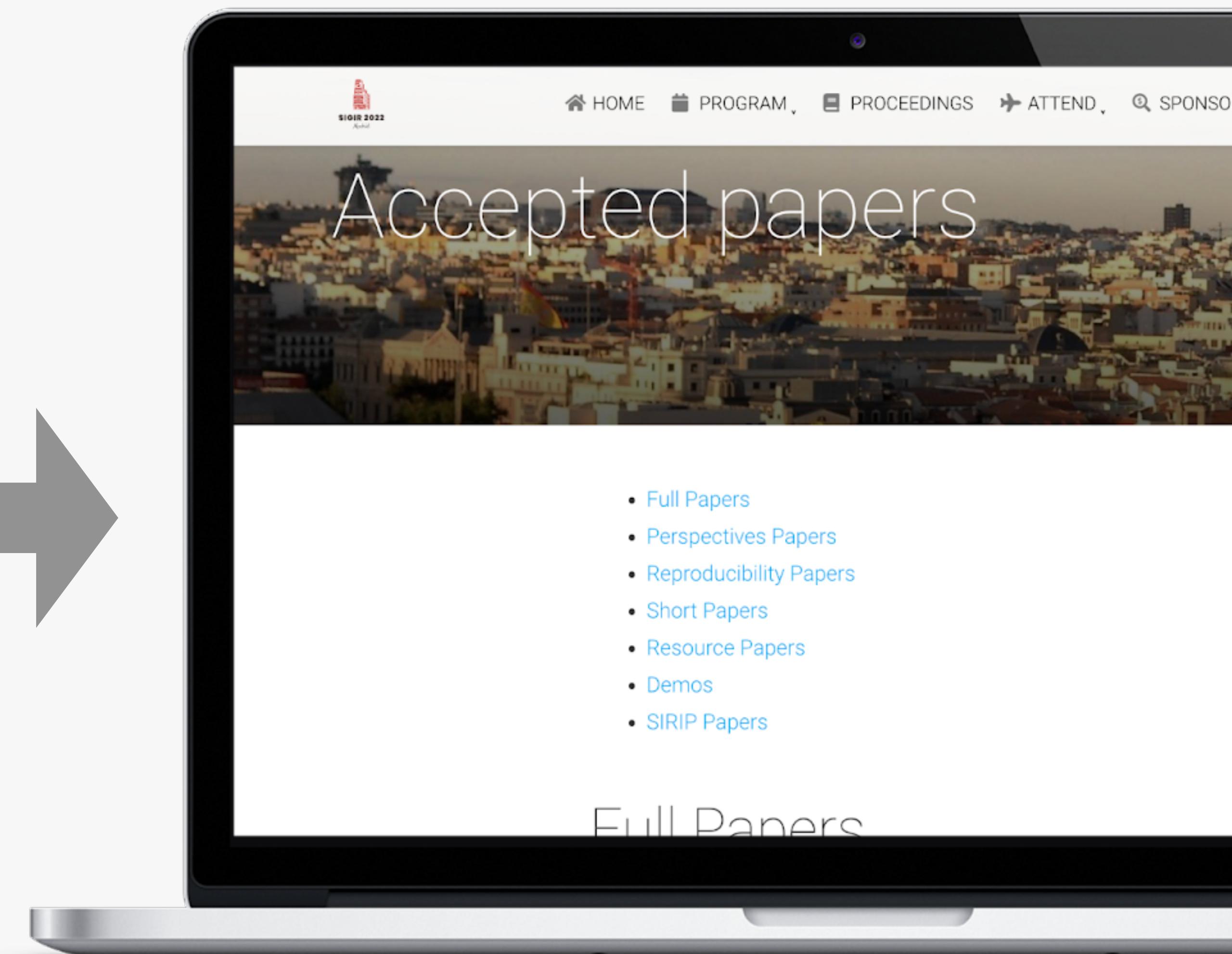
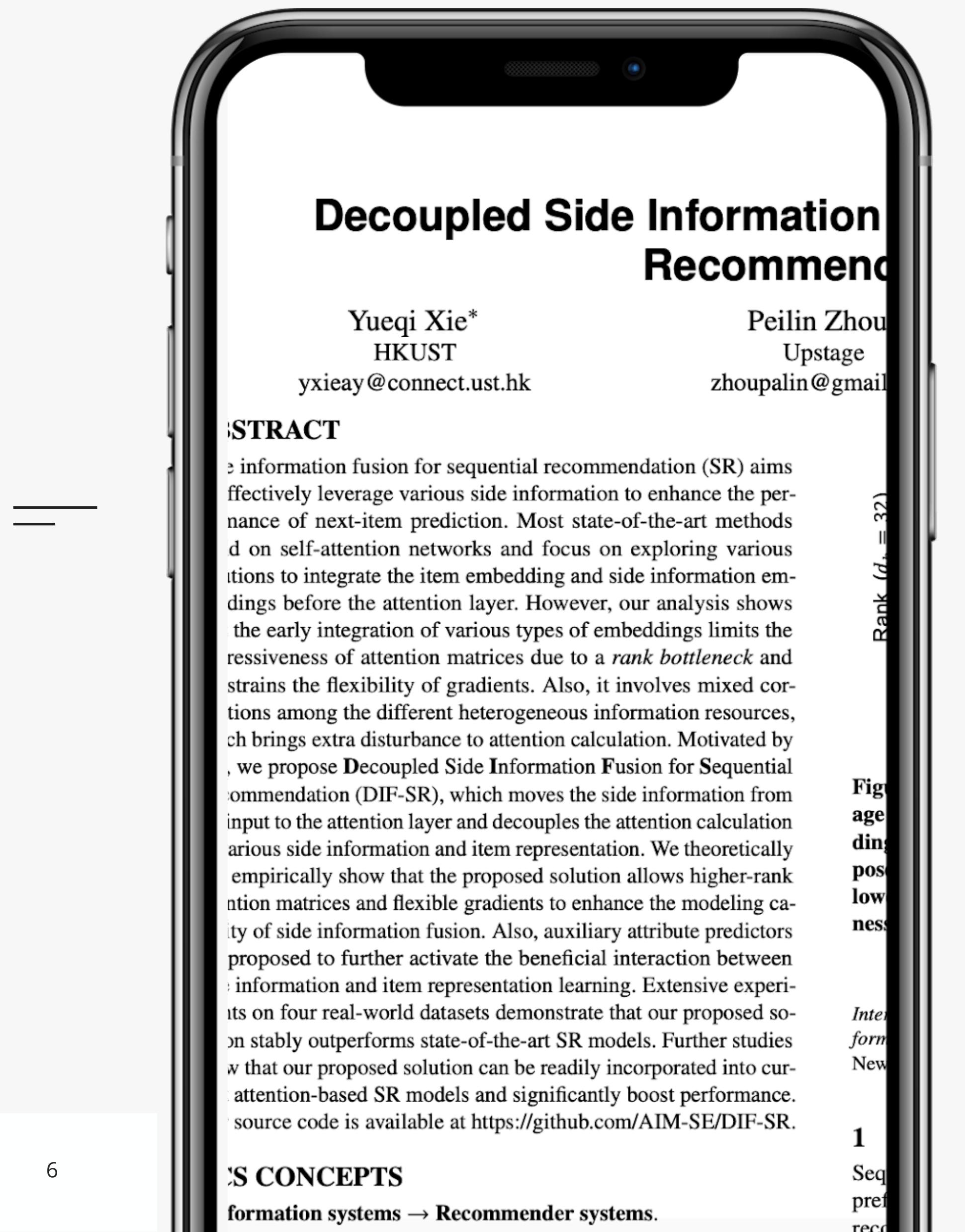
Problema Inicial



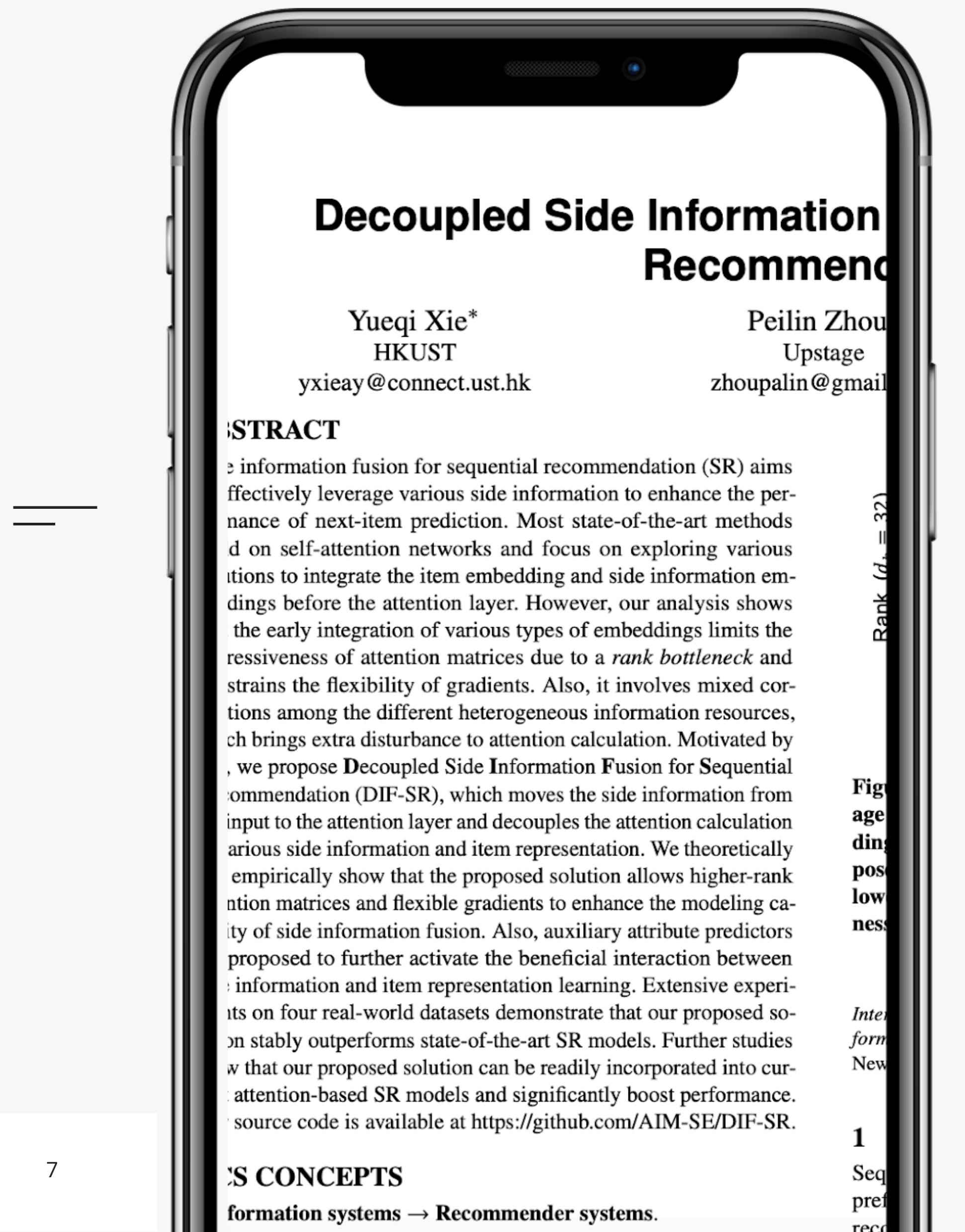
Problema Inicial



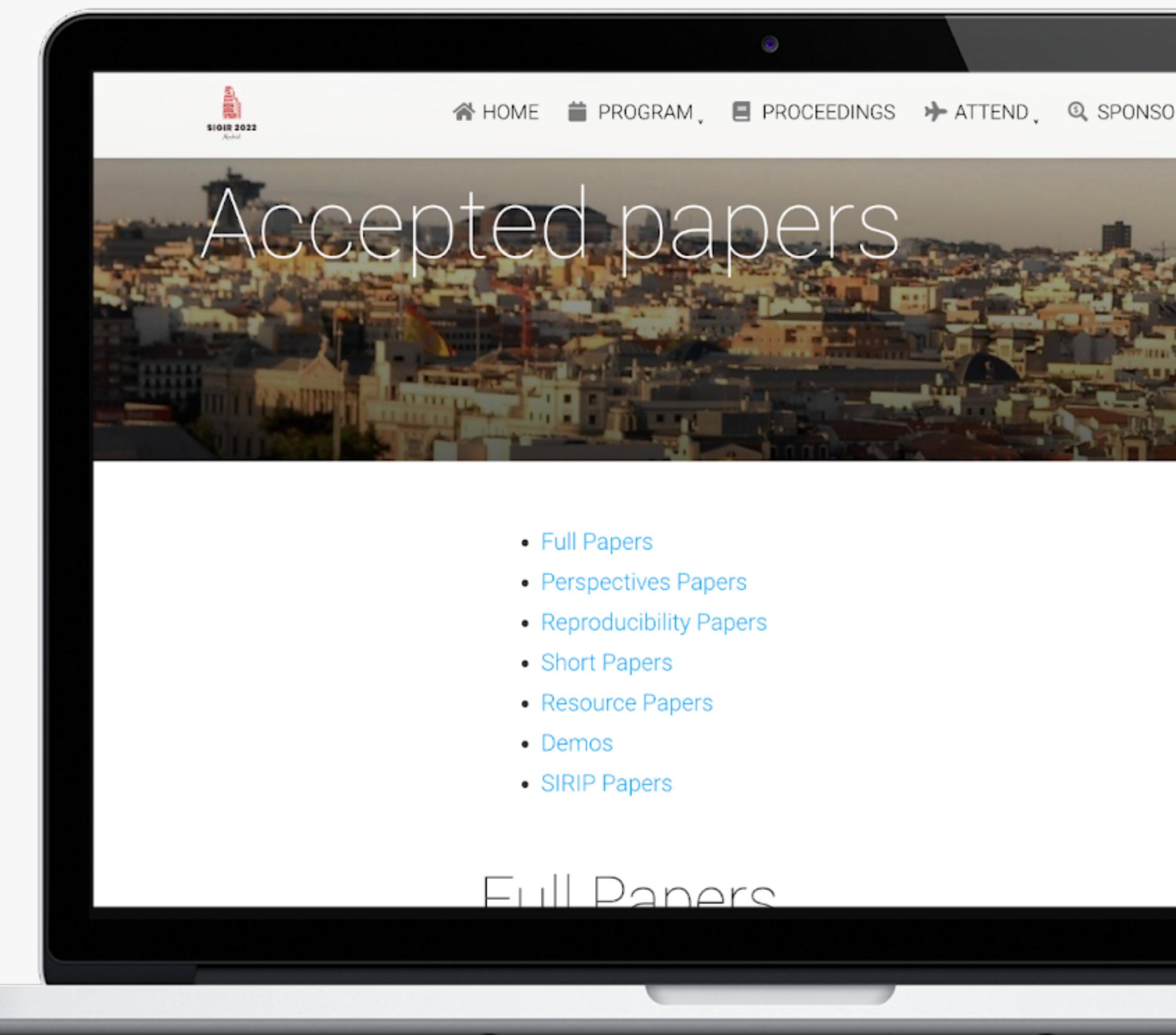
Problema Inicial



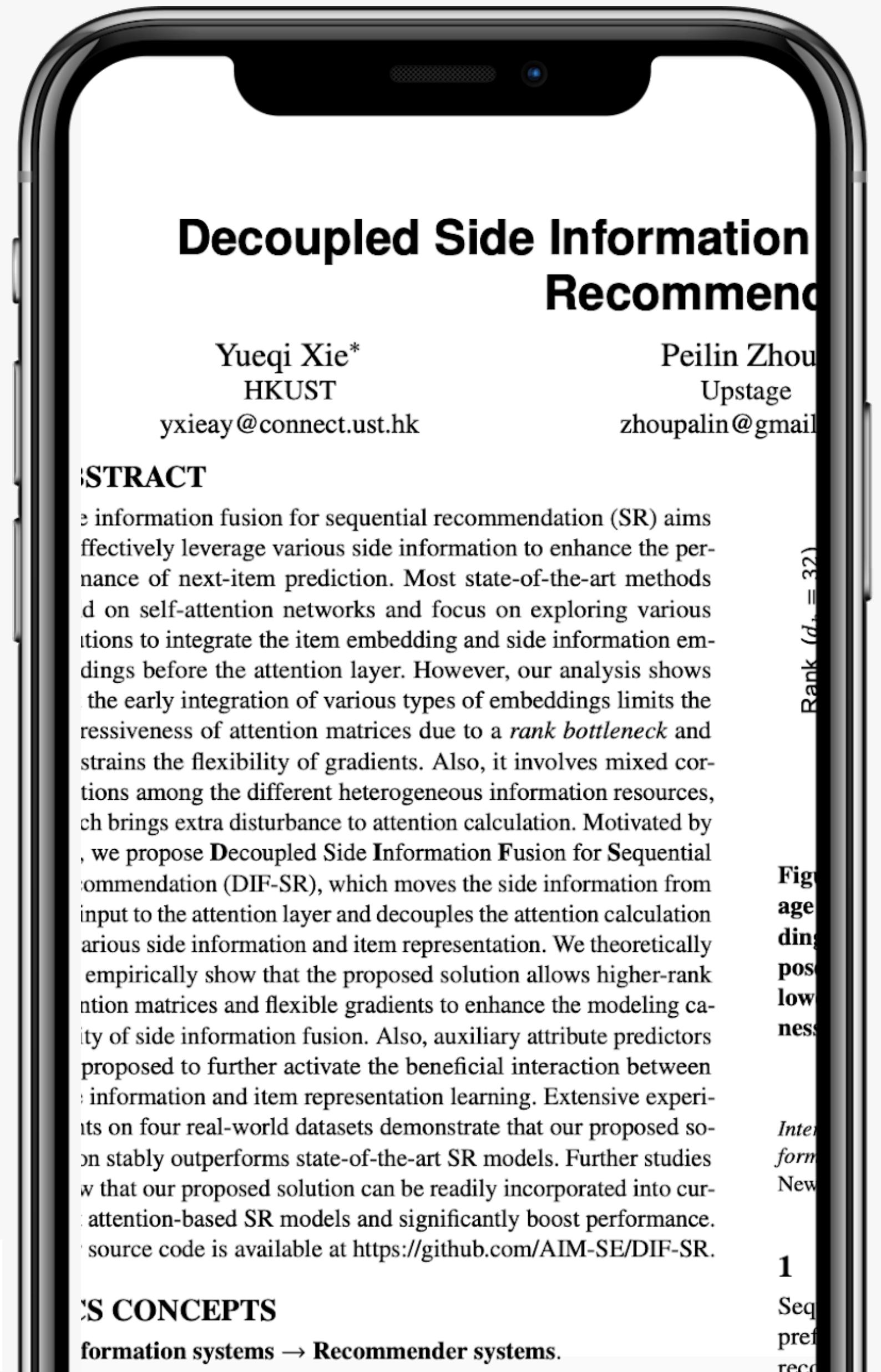
Problema Inicial



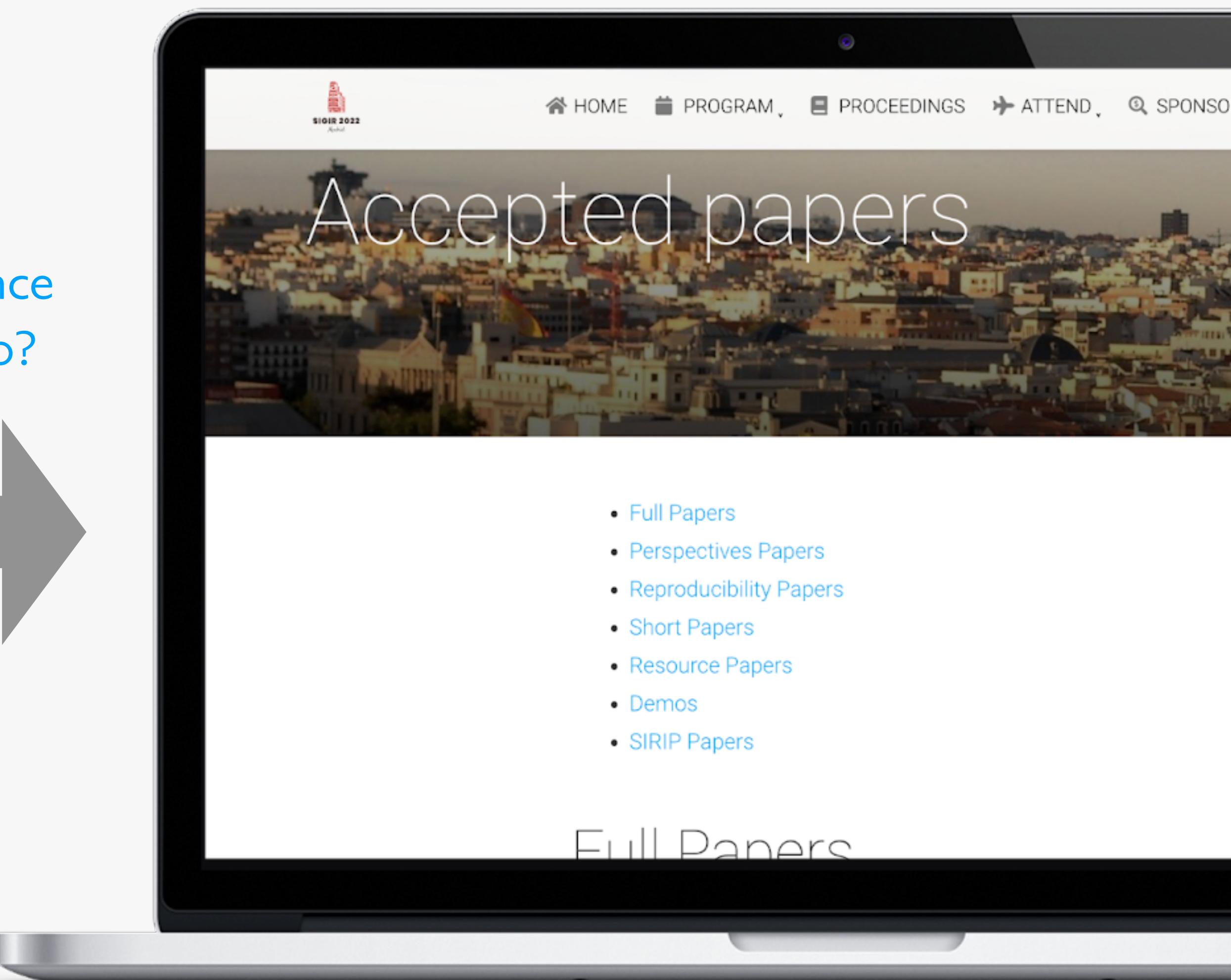
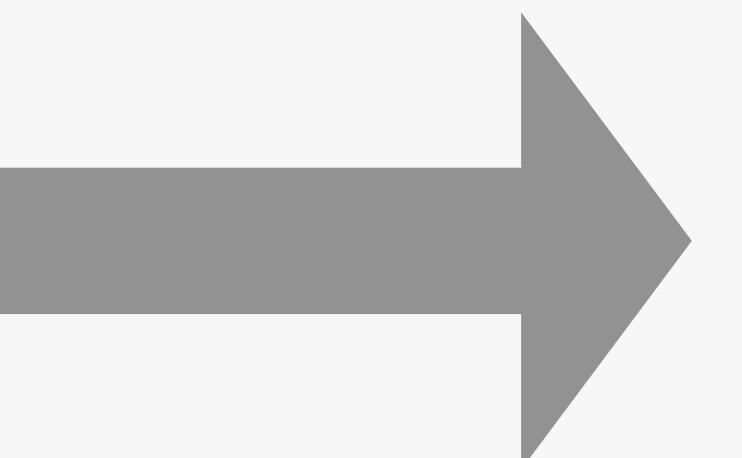
Ser aceito



Problema Inicial



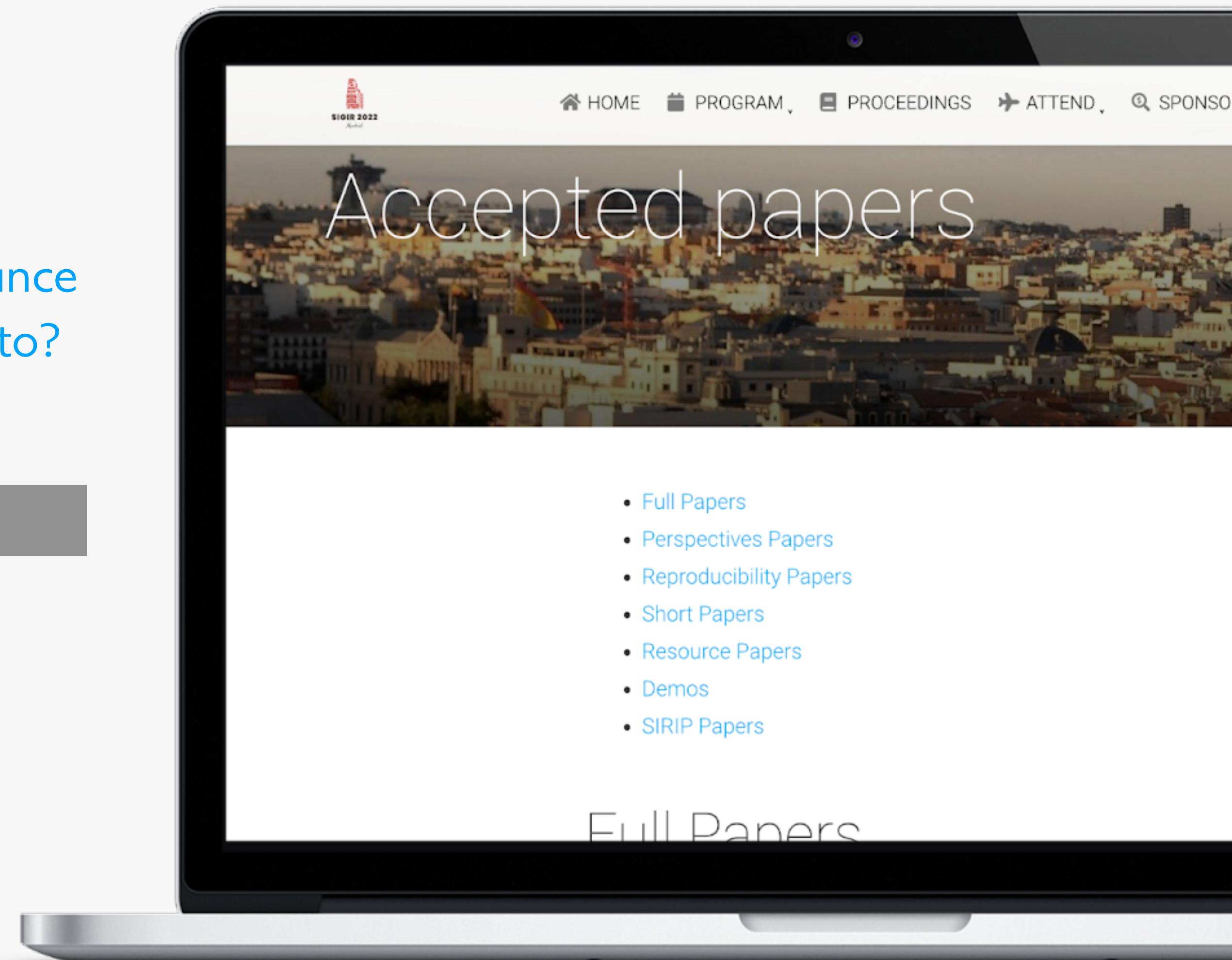
Qual a chance
de ser aceito?



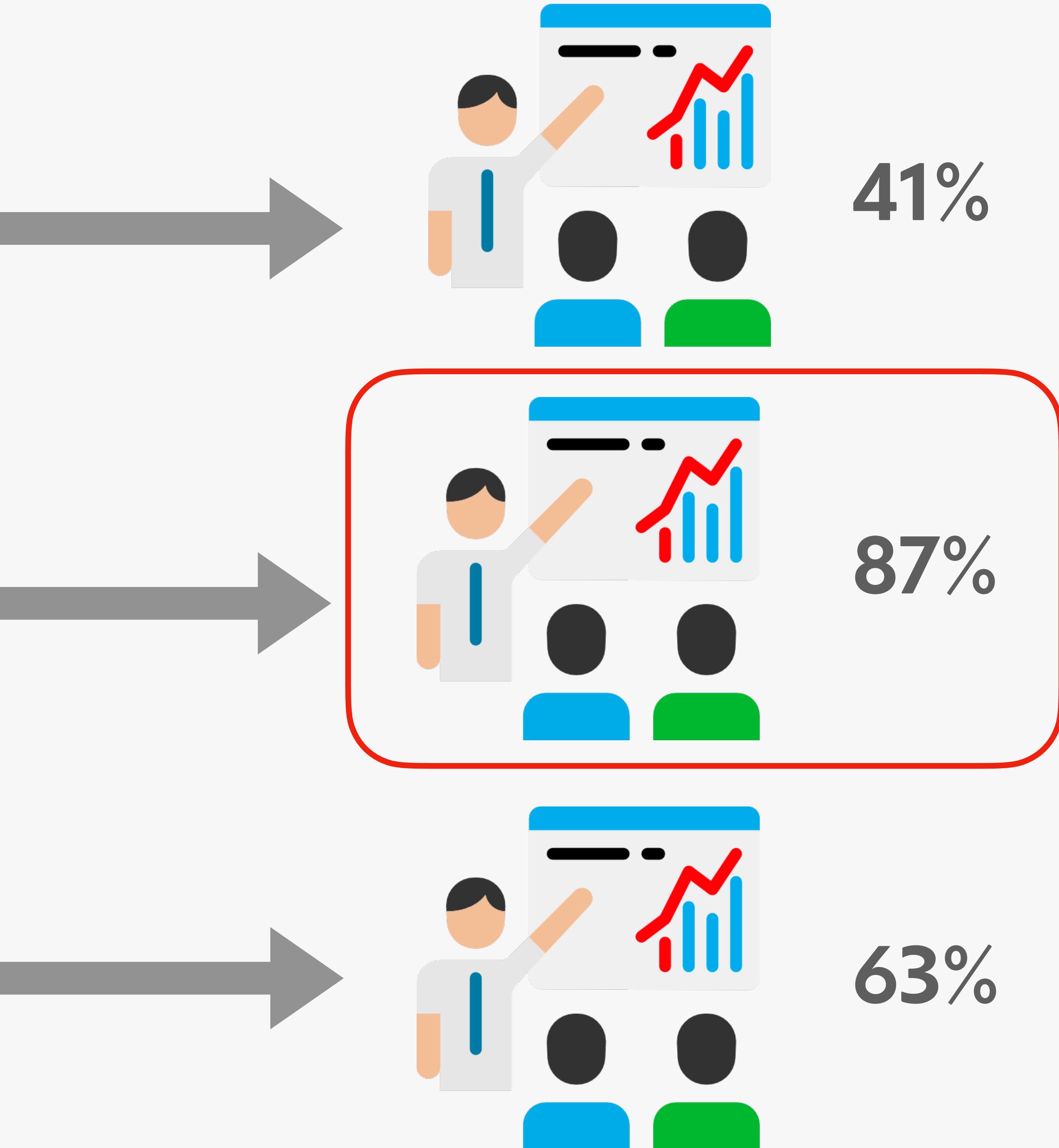
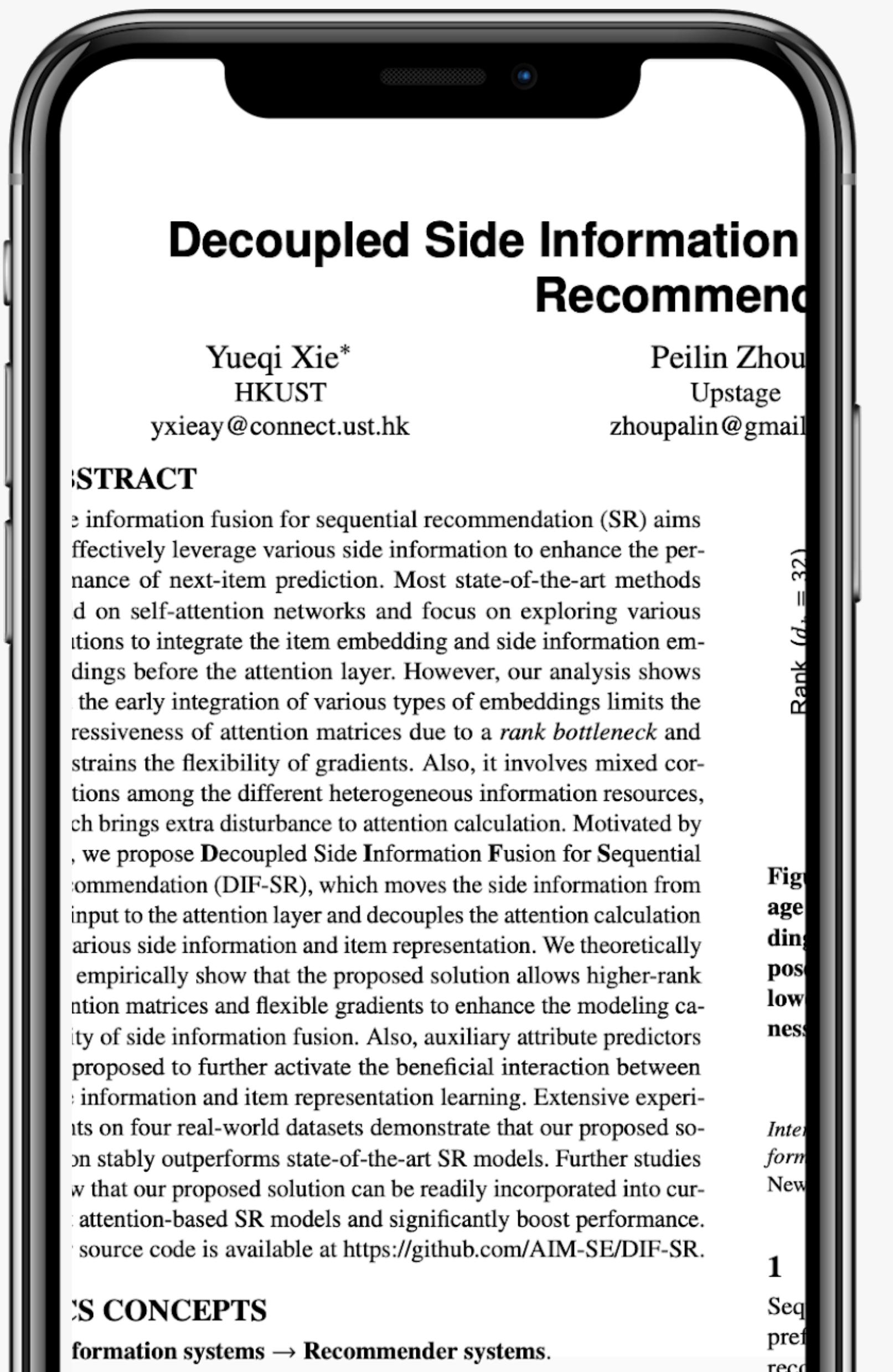
Problema Inicial



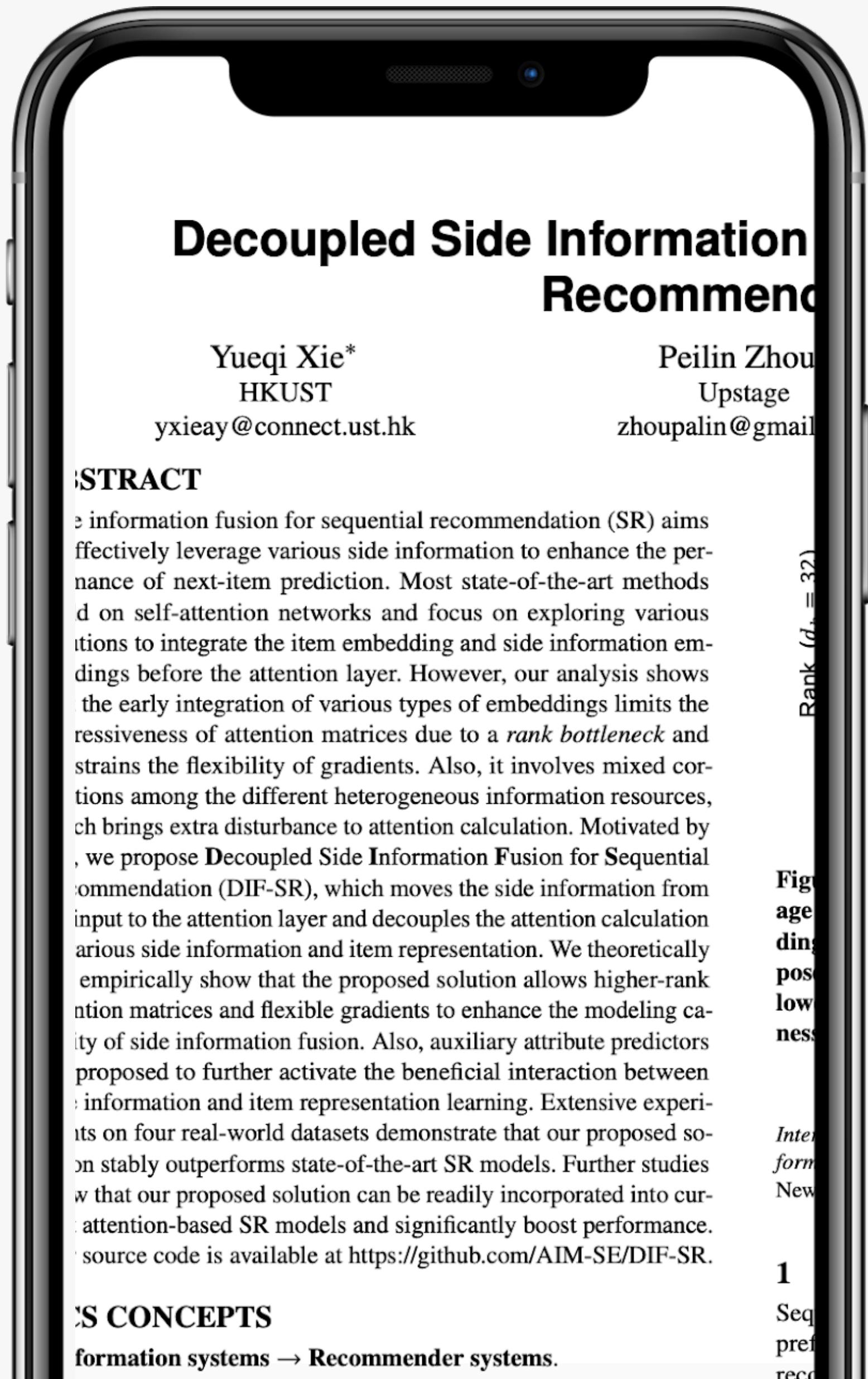
Qual a chance
de ser aceito?



Problema Inicial



Problema Inicial

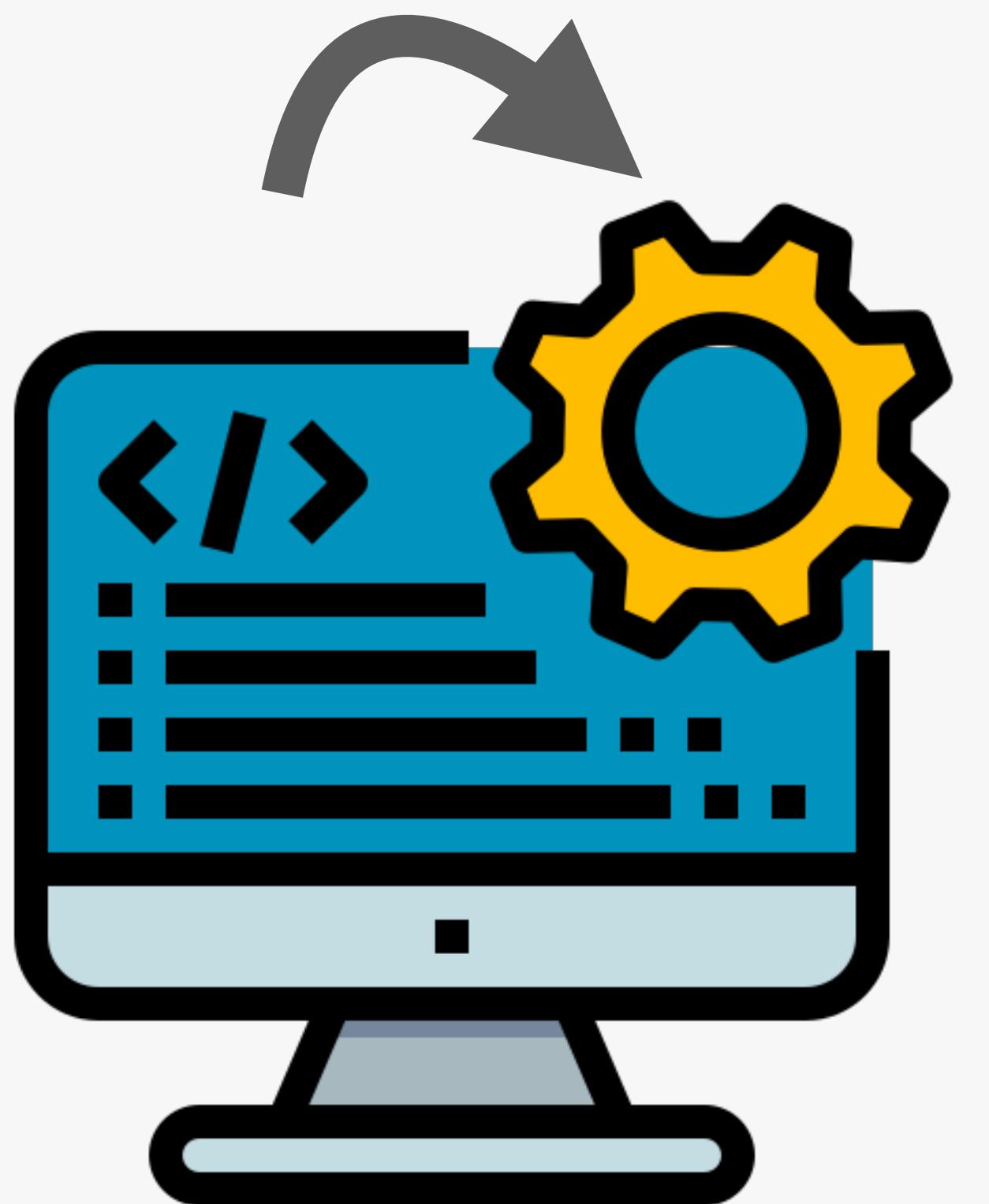


Faz o processamento utilizando dados de artigos já publicados em conferências

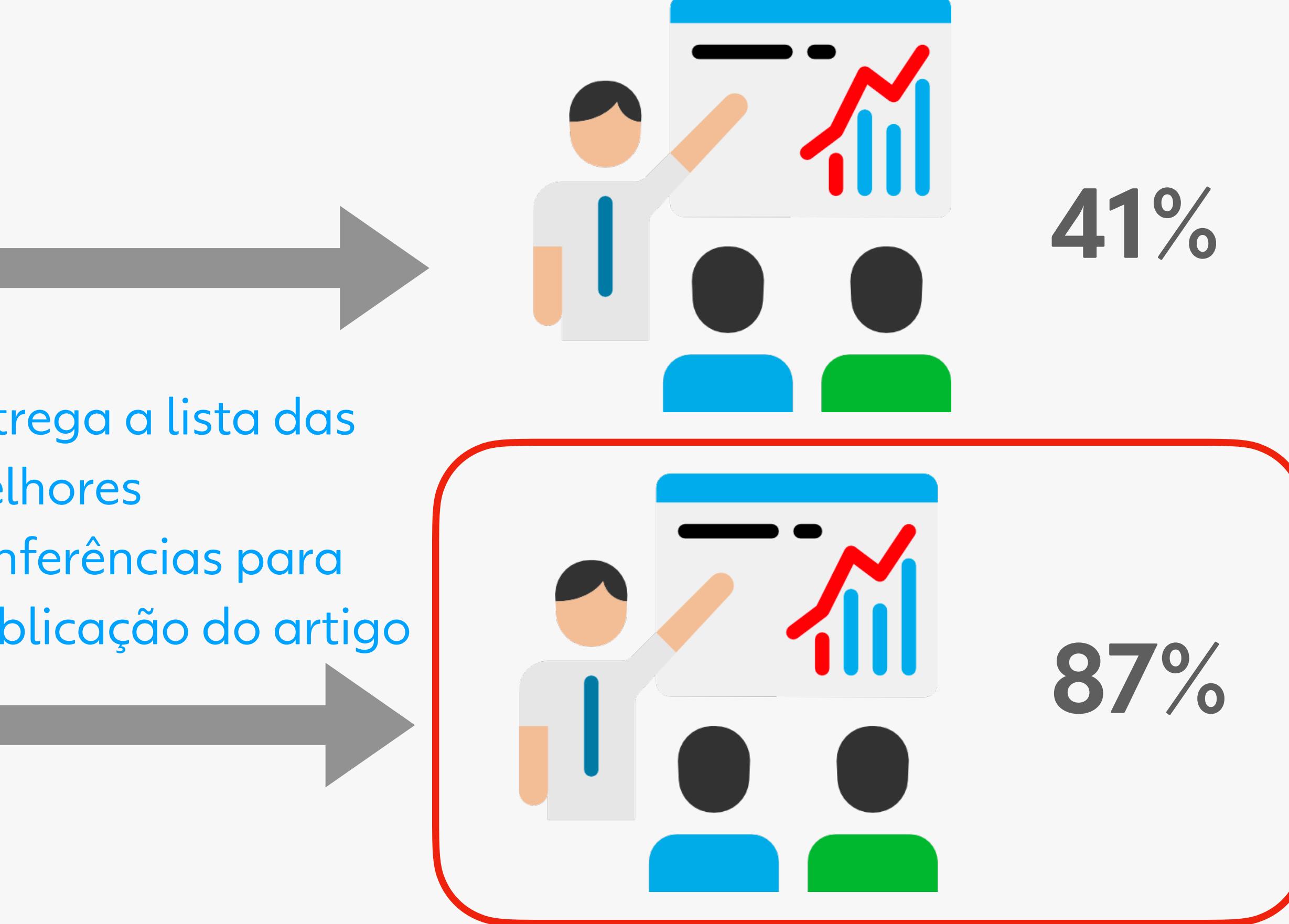


Problema Inicial

Faz o processamento utilizando dados de artigos já publicados em conferências

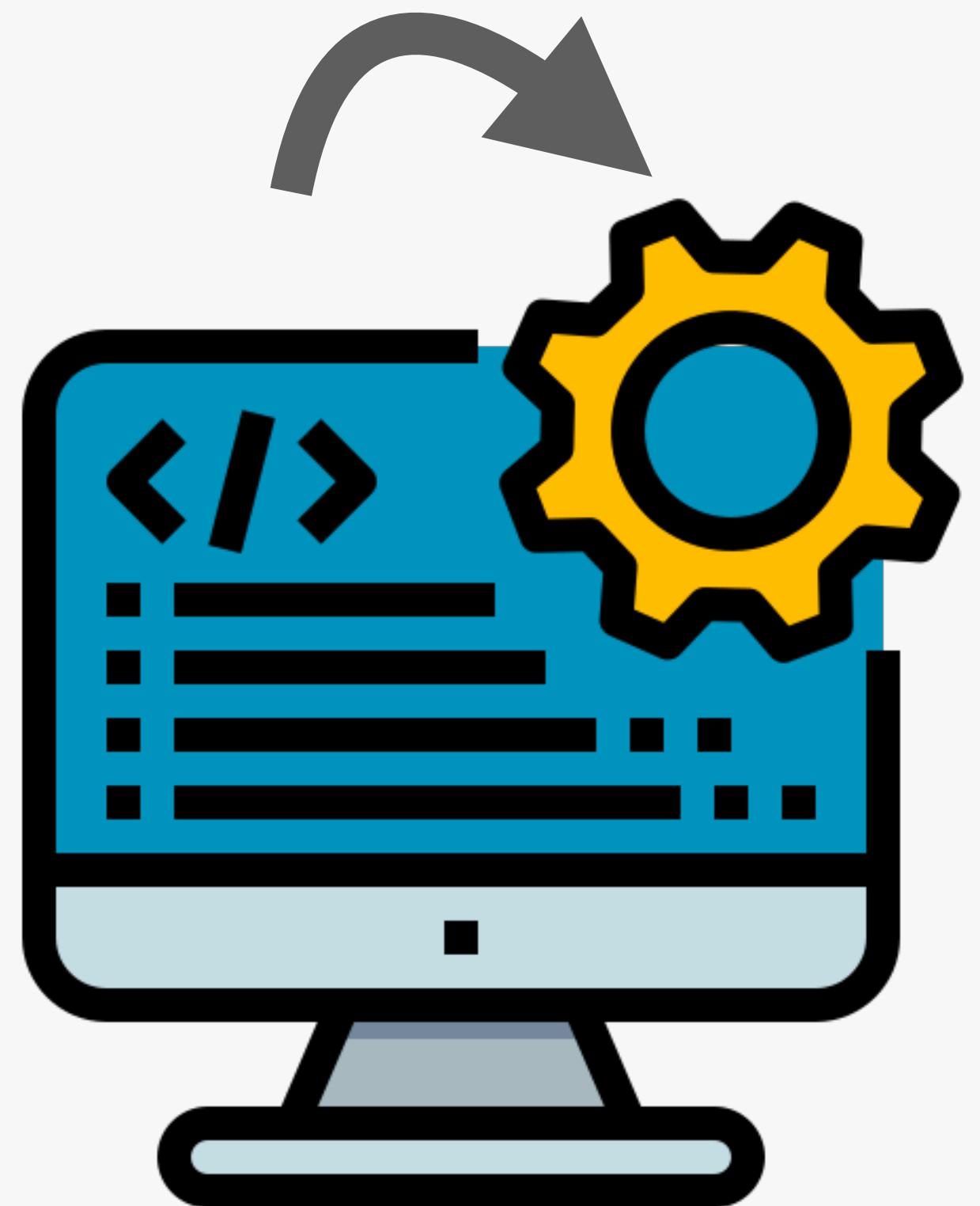


Entrega a lista das melhores conferências para publicação do artigo



Problema Inicial

Faz o processamento utilizando dados de artigos já publicados em conferências



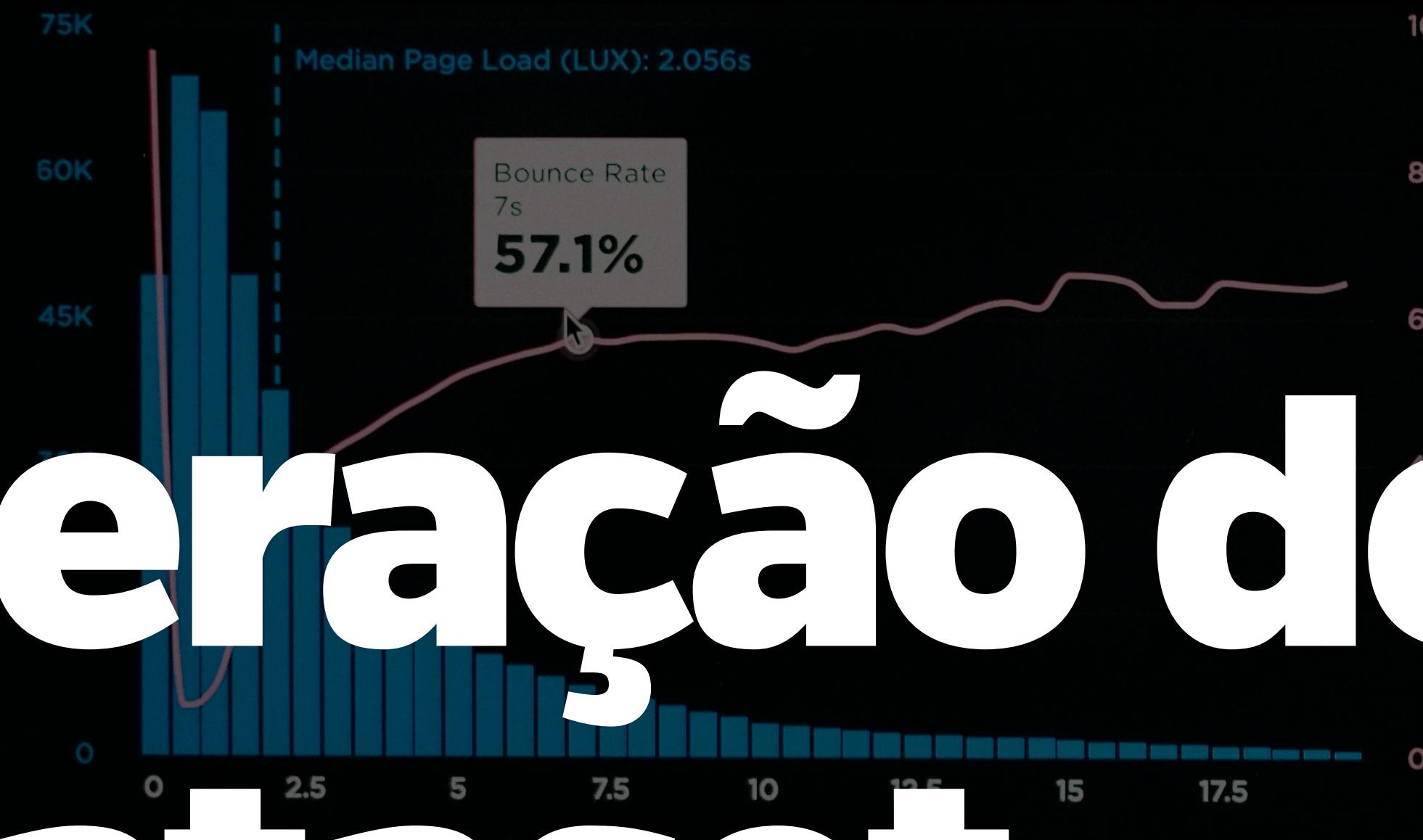
Entender como estão as conferências em relação aos temas publicados nelas e um histórico



USERS: LAST 7 DAYS USING MEDIAN ▾



LOAD TIME VS BOUNCE RATE



OPTIONS

100 %

80 %

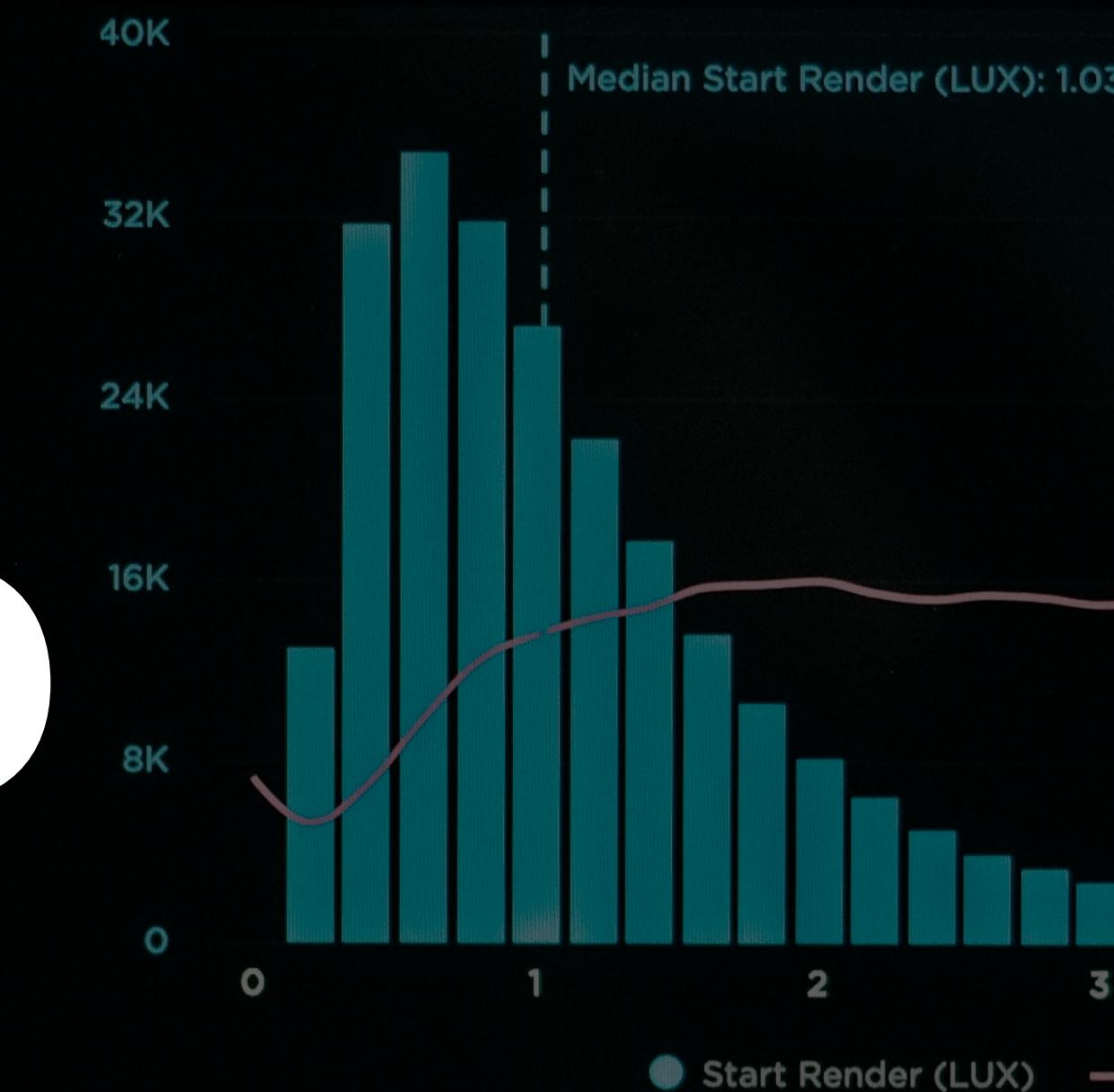
60 %

40 %

20 %

0 %

START RENDER VS BOUNCE RATE



OPTIONS

100 %

80 %

60 %

40 %

20 %

0 %

Geração do Dataset

Page Load (LUX)

0.7s

1s

Page Views (LUX)

2.7MpvS

1s

Bounce Rate (LUX)

40.6%

1s

SESSIONS

Sessions (LUX)

479K

4 pvs

PVs Per Session (LUX)

Session Length (LUX)

17min

100K 40 min

2pvs

80K 32 min

60K 24 min

40K



0.8s

0.6s

0.4s

500K 100%

400K 80%

300K 60%

200K 40%

3.2 pvs

2.4 pvs

1.6 pvs

Geração do dataset

10 conferências:

SOSP—ACM Symposium on Operating Systems Principles

OSDI—Operating Systems Design and Implementation

NDSS—Network and Distributed System Security Symposium

MobiHoc—Mobile Ad Hoc Networking and Computing

SIGCOMM—ACM SIGCOMM Conference

SenSys—Conference On Embedded Networked Sensor Systems

MOBICOM—Mobile Computing and Networking

CIDR—Conference on Innovative Data Systems Research

USENIX Security Symposium

EUROCRYPT—Theory and Application of Cryptographic Techniques

The following are the journals and conferences in computer science that have published at least 100 papers (2003–2013), with at least 5 citations per paper. Microsoft Academic Search. [download table]

Rank	Avg citations per paper	Conference
1.	66.3	CSUR—ACM Computing Surveys
2.	53.5	SOSP—ACM Symposium on Operating Systems Principles
3.	52.8	OSDI—Operating Systems Design and Implementation
4.	43.4	NDSS—Network and Distributed System Security Symposium
5.	38.7	MobiHoc—Mobile Ad Hoc Networking and Computing
6.	36.1	SIGCOMM—ACM SIGCOMM Conference
7.	35.3	SenSys—Conference On Embedded Networked Sensor Systems
8.	35.3	MOBICOM—Mobile Computing and Networking
9.	35.0	CIDR—Conference on Innovative Data Systems Research
10.	35.0	USENIX Security Symposium
11.	34.4	EUROCRYPT—Theory and Application of Cryptographic Techniques
12.	33.5	NSDI—Networked Systems Design and Implementation
13.	33.5	JASSS—The Journal of Artificial Societies and Social Simulation
14.	33.4	TOCS—ACM Transactions on Computer Systems
15.	33.4	S&P—IEEE Symposium on Security and Privacy
16.	33.4	MobiSys—International Conference on Mobile Systems
17.	32.5	IJCV—International Journal of Computer Vision
18.	32.2	TOG—ACM Transactions on Graphics/SIGGRAPH
19.	31.6	VLDB—Very Large Data Bases
20.	30.9	BioMED—Biomedical Engineering
21.	30.9	IEEE TRANS ROBOTICS AUTOMAT—IEEE Transactions on Robotics and Automation
22.	30.6	CRYPTO—International Cryptology Conference
23.	30.1	PAMI—IEEE Transactions on Pattern Analysis and Machine Intelligence
24.	29.6	PLDI—SIGPLAN Conference on Programming Language Design and Implementation
25.	29.3	MICRO—International Symposium on Microarchitecture
26.	29.1	Journal of Web Semantics
27.	28.5	BIB—Briefings in Bioinformatics
28.	27.4	JMLR—Journal of Machine Learning Research
29.	27.3	ICML—International Conference on Machine Learning

Geração do dataset

10 conferências:

SOSP—ACM Symposium on Operating Systems Principles

OSDI—Operating Systems Design and Implementation

NDSS—Network and Distributed System Security Symposium

MobiHoc—Mobile Ad Hoc Networking and Computing

SIGCOMM—ACM SIGCOMM Conference

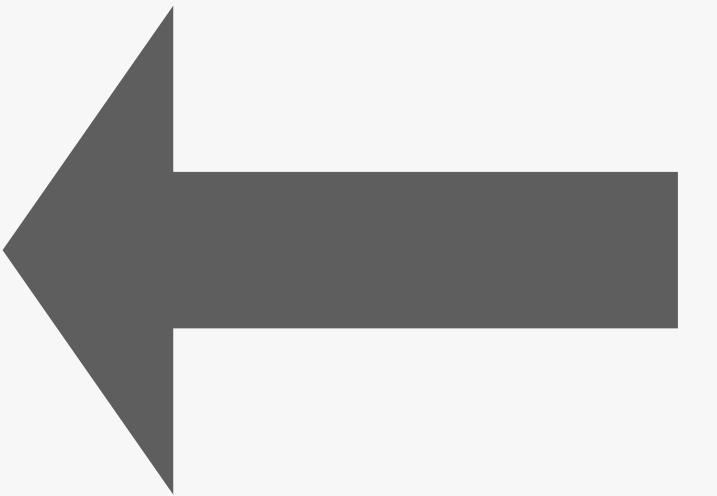
SenSys—Conference On Embedded Networked Sensor Systems

MOBICOM—Mobile Computing and Networking

CIDR—Conference on Innovative Data Systems Research

USENIX Security Symposium

EUROCRYPT—Theory and Application of Cryptographic Techniques



2018 a 2022

Geração do dataset

10 conferências:

SOSP—ACM Symposium on Operating Systems Principles

OSDI—Operating Systems Design and Implementation

NDSS—Network and Distributed System Security Symposium

MobiHoc—Mobile Ad Hoc Networking and Computing

SIGCOMM—ACM SIGCOMM Conference

SenSys—Conference On Embedded Networked Sensor Systems

MOBICOM—Mobile Computing and Networking

CIDR—Conference on Innovative Data Systems Research

USENIX Security Symposium

EUROCRYPT—Theory and Application of Cryptographic Techniques



Geração do dataset

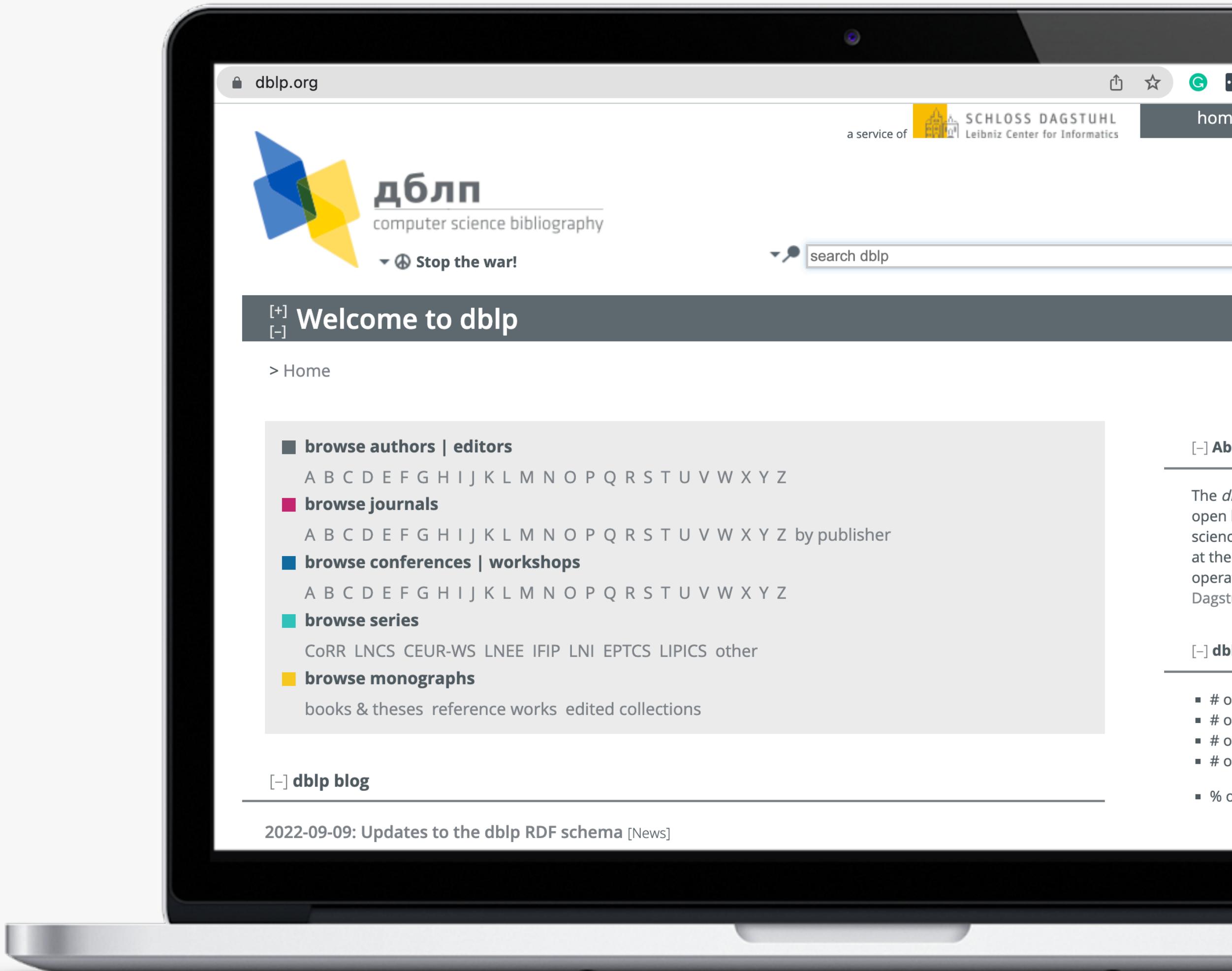
Artigos são extraídos do site da **Dblp**

[publications](#): 6,303,209

[authors](#): 3,085,450

[conferences](#): 5,978

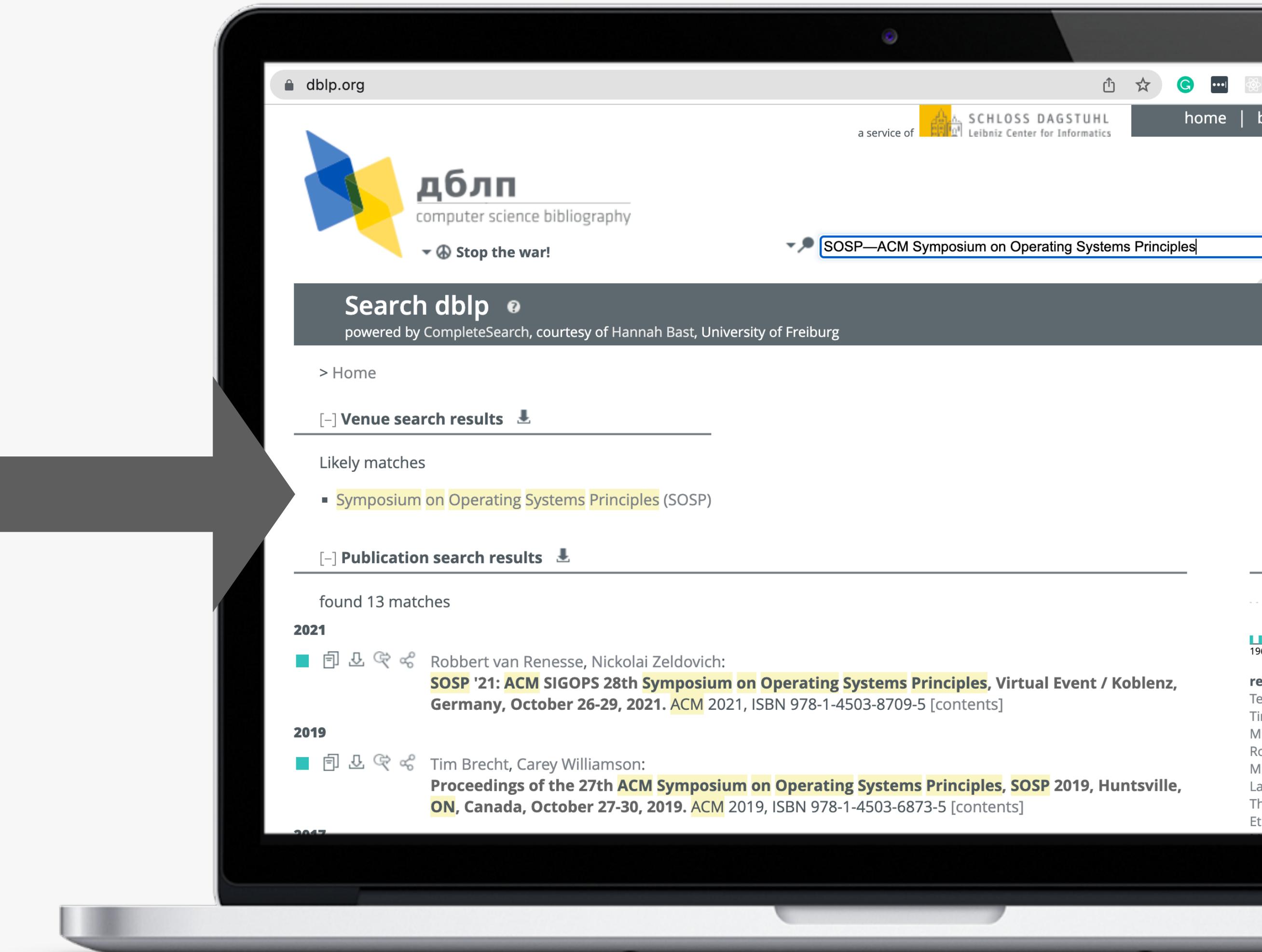
[journals](#): 1,815



Geração do dataset

Para cada conferência,
ela é buscada na rota
de busca do site

(<https://dblp.org/search?q=%>)



Geração do dataset

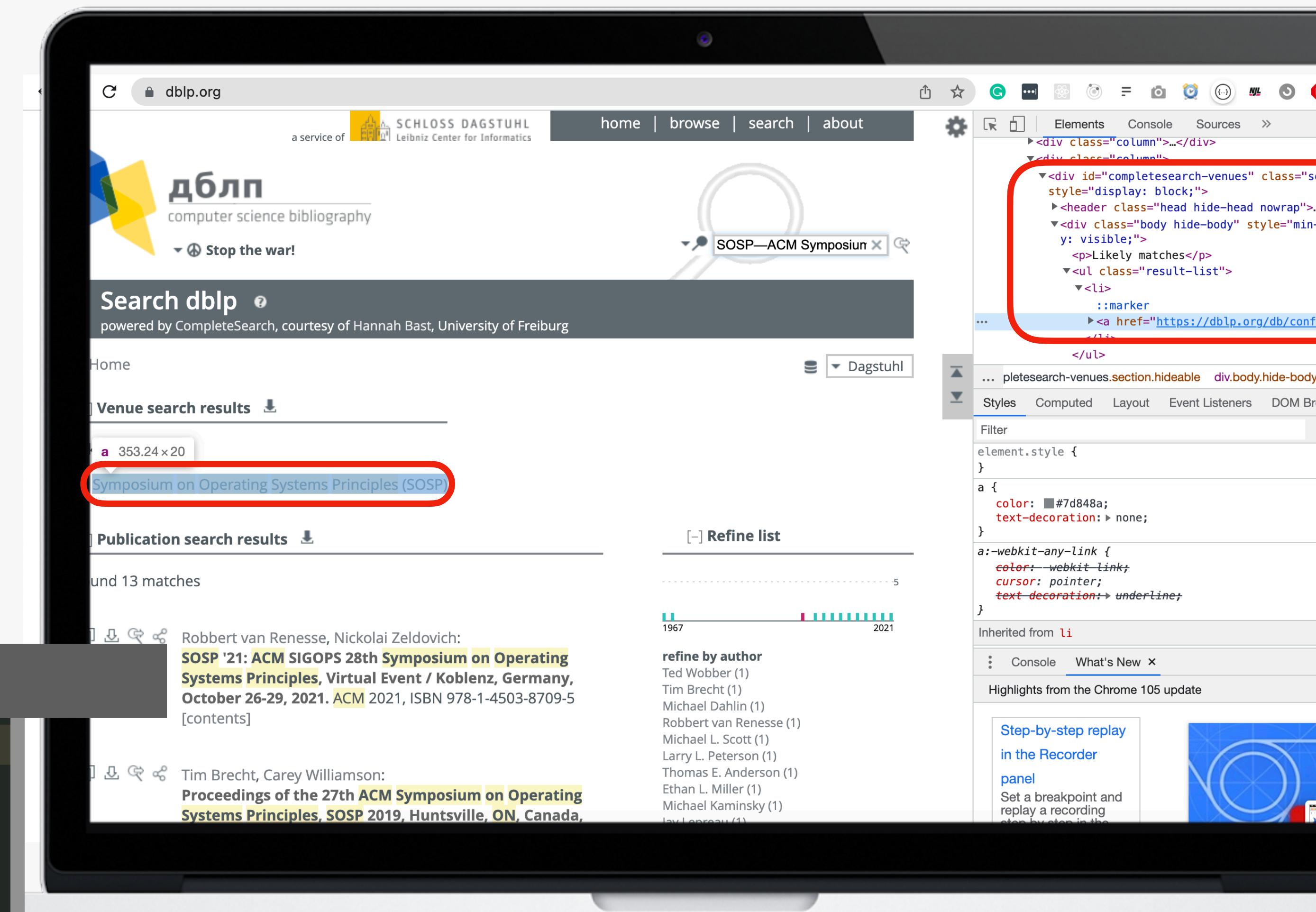
Utiliza Web Scraper
para entrar no link de
listagem da
conferência

```
# Formatando o retorno
sopa = BeautifulSoup(resposta.text, 'html.parser')

infos = sopa.find('div', {'id':'completesearch-venues'})

venue = infos.find('ul', {'class':'result-list'}).find('a')

print('Href: ' + venue.get('href'))
venue_href = venue.get('href')
resposta = requests.get(venue_href, headers = cabecalho)
```



Geração do dataset

A partir da listagem de edições, entra no link das que são entre 2018 e 2022

```
for conf_edition in sopa.find_all('ul', {'class':'publ-list'}):
    conf_title = conf_edition.find('span',{'itemprop':'name'}).get_text().lower()

    # Desconsidera os workshops
    if 'workshop' in conf_title:
        continue

    published_year = conf_edition.find('span',{'itemprop':'datePublished'}).get_text()
    print('Ano: ' + published_year)

    # Só pega os artigos ate 2018
    if int(published_year) < 2018:
        break
    content = conf_edition.find('a',{'class':'toc-link'}).get('href')
    print(content)
    resposta = requests.get(content, headers = cabecalho)
```

dblp.org/db/conf/sosp/index.html

2021: Virtual Event / Koblenz, Germany

Robbert van Renesse, Nickolai Zeldovich:
SOSP '21: ACM SIGOPS 28th Symposium on Operating Systems Principles, Virtual Event / Koblenz, Germany, October 25-29, 2021. ACM 2021, ISBN 978-1-4503-8709-5

[\[contents\]](#)

PLOS '21: Proceedings of the 11th Workshop on Programming Languages and Operating Systems, Virtual Event, Germany, October 25, 2021. ACM 2021, ISBN 978-1-4503-8707-1 [contents]

ResilientFL '21: Proceedings of the First Workshop on Systems Challenges in Reliable and Secure Federated Learning, Virtual Event / Koblenz, Germany, 25 October 2021. ACM 2021, ISBN 978-1-4503-8708-8 [contents]

2019: Huntsville, ON, Canada

Tim Brecht, Carey Williamson:
Proceedings of the 27th ACM Symposium on Operating Systems Principles, SOSP 2019, Huntsville, ON, Canada, October 27-30, 2019. ACM 2019, ISBN 978-1-4503-6873-5 [contents]

Proceedings of the 10th Workshop on Programming Languages and Operating Systems, SOSP 2019, Huntsville, ON, Canada, October 27-30, 2019. ACM 2019, ISBN 978-1-4503-7017-2 [contents]

Proceedings of the 4th Workshop on System Software for Trusted Execution, SysTEX@SOSP 2019, Huntsville, ON, Canada, October 27-30, 2019. ACM 2019, ISBN 978-1-4503-6888-9 [contents]

2017: Shanghai, China

Proceedings of the 26th Symposium on Operating Systems Principles, Shanghai, China, October 28-31, 2017. ACM 2017, ISBN 978-1-4503-5085-3 [contents]

Julia Lawall:

Roger M. Needham (13)
Barbara Liskov (13)
Andrea C. Arpacı-Dusseau (11)
Remzi H. Arpacı-Dusseau (11)
Gerald J. Popek (10)
Yuanyuan Zhou 0001 (10)
Mahadev Satyanarayanan (10)

show more

Elements Console Sources > 1

...
", "
...
": "

...
ACM
2021
", ISBN "
span itemprop="isbn">978-1-4503-8709-5
[contents] == \$0
... ul.publ-list li#confVsospV2021.entry.editor.toc cite.data.tts-content
...
Styles Computed Layout Event Listeners DOM Breakpoints

:hov .cls dblp-2022- user agent
element.style {
}
a {
color: #7d848a;
text-decoration: none;
}
a:-webkit-any-link {
color: webkit-link;
cursor: pointer;
text-decoration: underline;
}
Inherited from cite.data.tts-content
Console What's New x
Highlights from the Chrome 105 update
Step-by-step replay in the Recorder panel
Set a breakpoint and replay a recording step by step in the Recorder panel

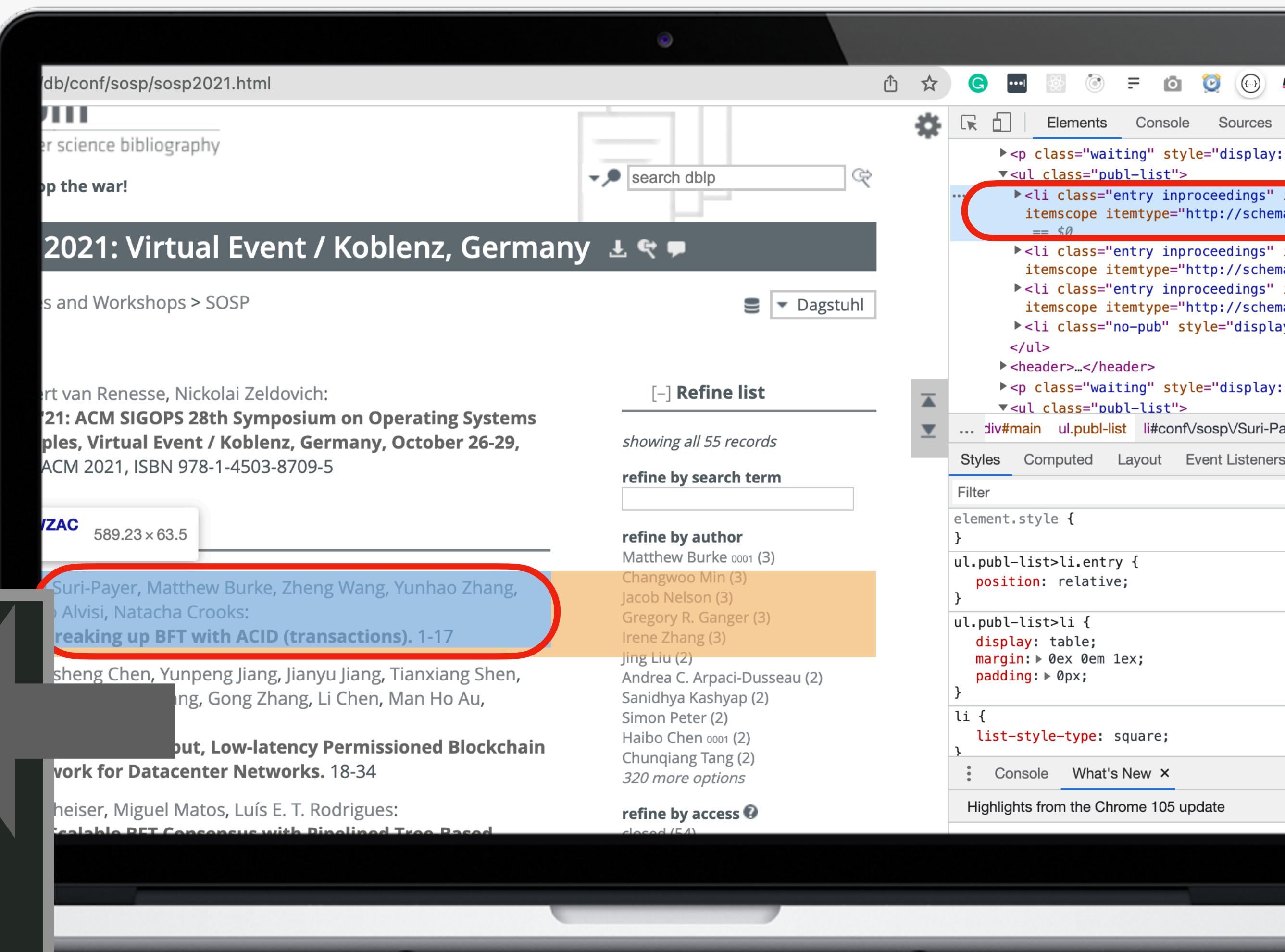
Geração do dataset

Entra em cada artigo,
pegando os dados dele:
**título, ano, gênero (CC), doi,
conferência**

```
for article in sopa.find_all('li', {'class':'entry inproceedings'}):
    header = article.find_all_previous('h2')[0].get('id').lower()

    # Desconsidera os keynotes
    if 'keynote' in header:
        continue

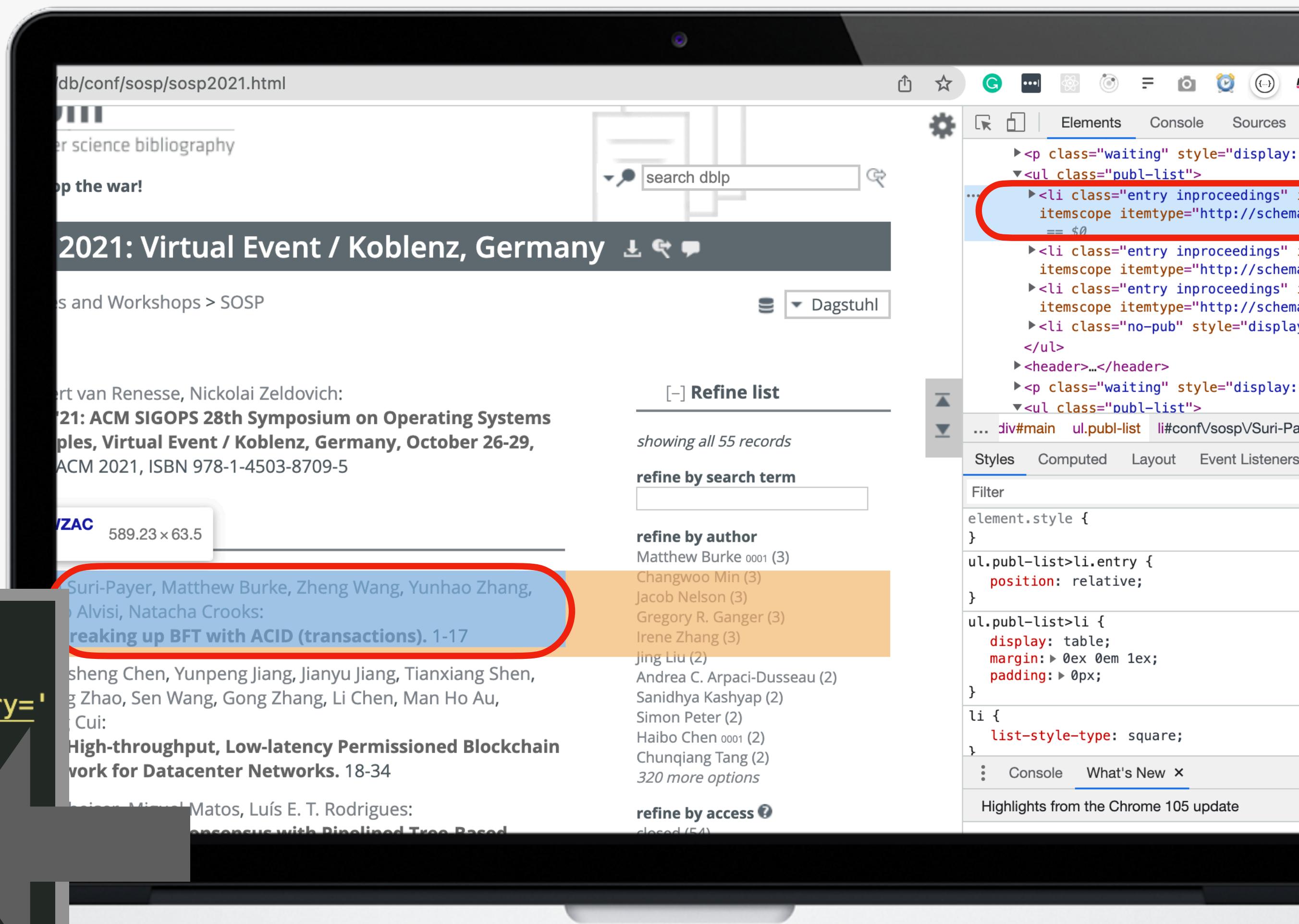
    title = article.find('span', {'class':'title'}).get_text()
    print(title)
    year = article.find('meta', {'itemprop':'datePublished'})['content']
    print(year)
    genre = article.find('meta', {'property':'genre'})['content']
    print(genre)
    doi_url = article.find('li', {'class': 'drop-down'}).find('a').get('href')
```



Geração do dataset

Quando não tem doi,
faz uma busca na api
do Semantic Scholar

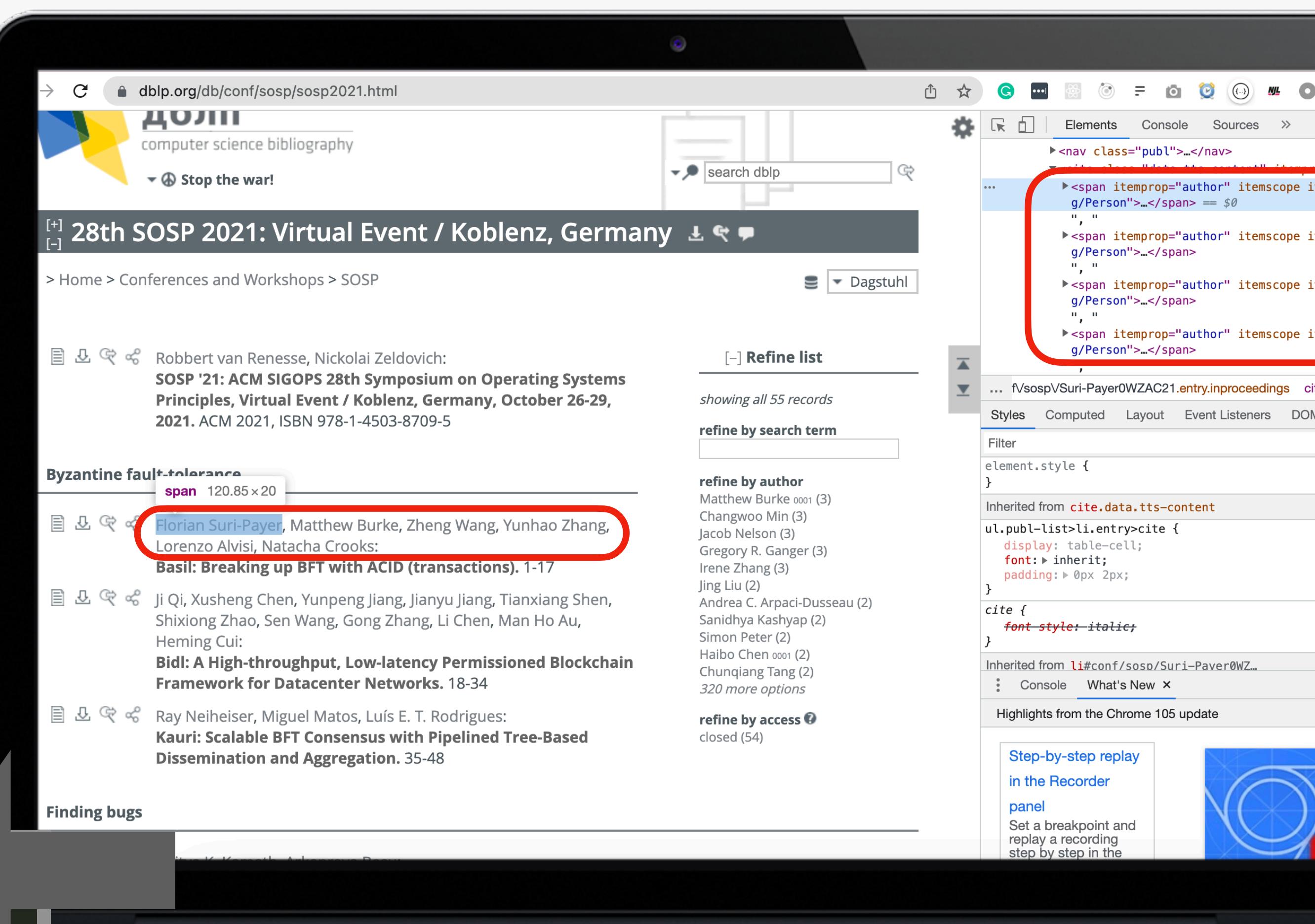
```
def findArticle(title, cabecalho):
    filtro = title.replace(' ', '+')
    link = 'https://api.semanticscholar.org/graph/v1/paper/search?query='
    print('Não tem DOI')
    retorno = requests.get(link + filtro, headers = cabecalho)
    if retorno.status_code == 429:
        time.sleep(300)
        print('Esperando um pouco')
        retorno = requests.get(link + filtro, headers = cabecalho)
    sp_artigo = BeautifulSoup(retorno.text, 'html.parser')
    artigo_json = json.loads(sp_artigo.text)
    if len(artigo_json["data"]) > 0:
        str_paperId = artigo_json["data"][0]["paperId"]
        return str_paperId
    return None
```



Geração do dataset

Se não tem **paperId**,
pega os autores do
próprio artigo e cria
uma estrutura básica

```
if paperId is None:  
    print('paperId is None')  
    authors = ''  
  
    for author in article.find_all('span',{'itemprop':'author'}):  
        authors = authors + author.find('span',{'itemprop':'name'}).get('title') + ','  
    authors = authors[0:len(authors)-1]  
    article_data = ['', '', authors, conf, '', genre]
```

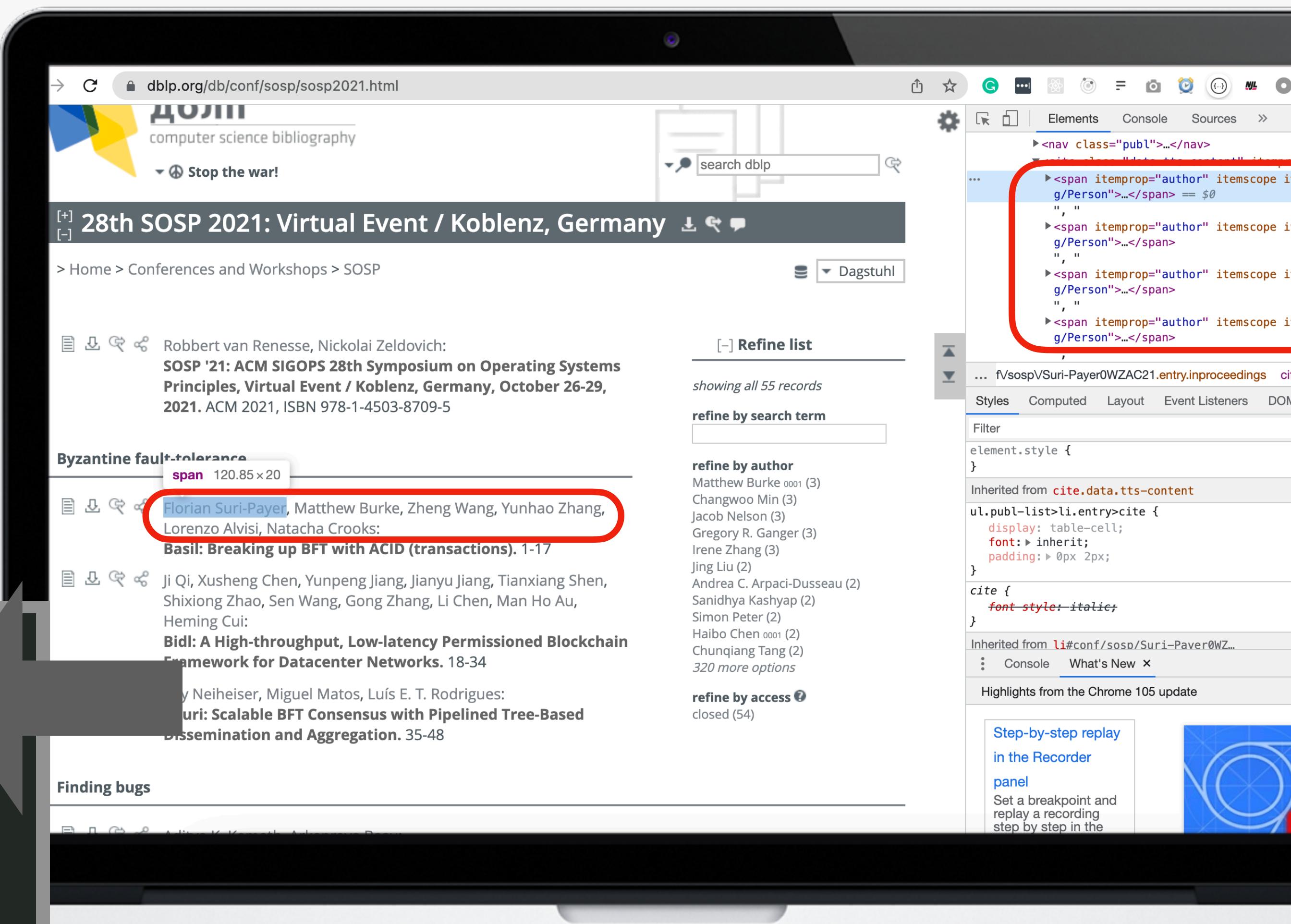


The screenshot shows a web browser displaying the dblp.org conference page for the 28th SOSP 2021. The page lists various research papers. In the developer tools element inspector, a red box highlights a section of the code where multiple `` elements are listed. This indicates that the script is extracting author information directly from the page's HTML structure.

Geração do dataset

Caso contrário,
complementa com dados da
API do semantic scholar
(lib do Python): **fieldsOfStudy**,
authors, **topics...**

```
else:  
    print('paperId: ' + paperId)  
    article_data = get_complementary_data(qty_article, paperId)  
    print('articledata1: ', article_data)  
    doi = article_data[1]  
    print('doi: ', doi)  
  
else:  
    doi = doi_url.split('https://doi.org/')[1]  
    print(doi)  
    article_data = get_complementary_data(qty_article, doi)  
    print('articledata2: ', article_data)
```



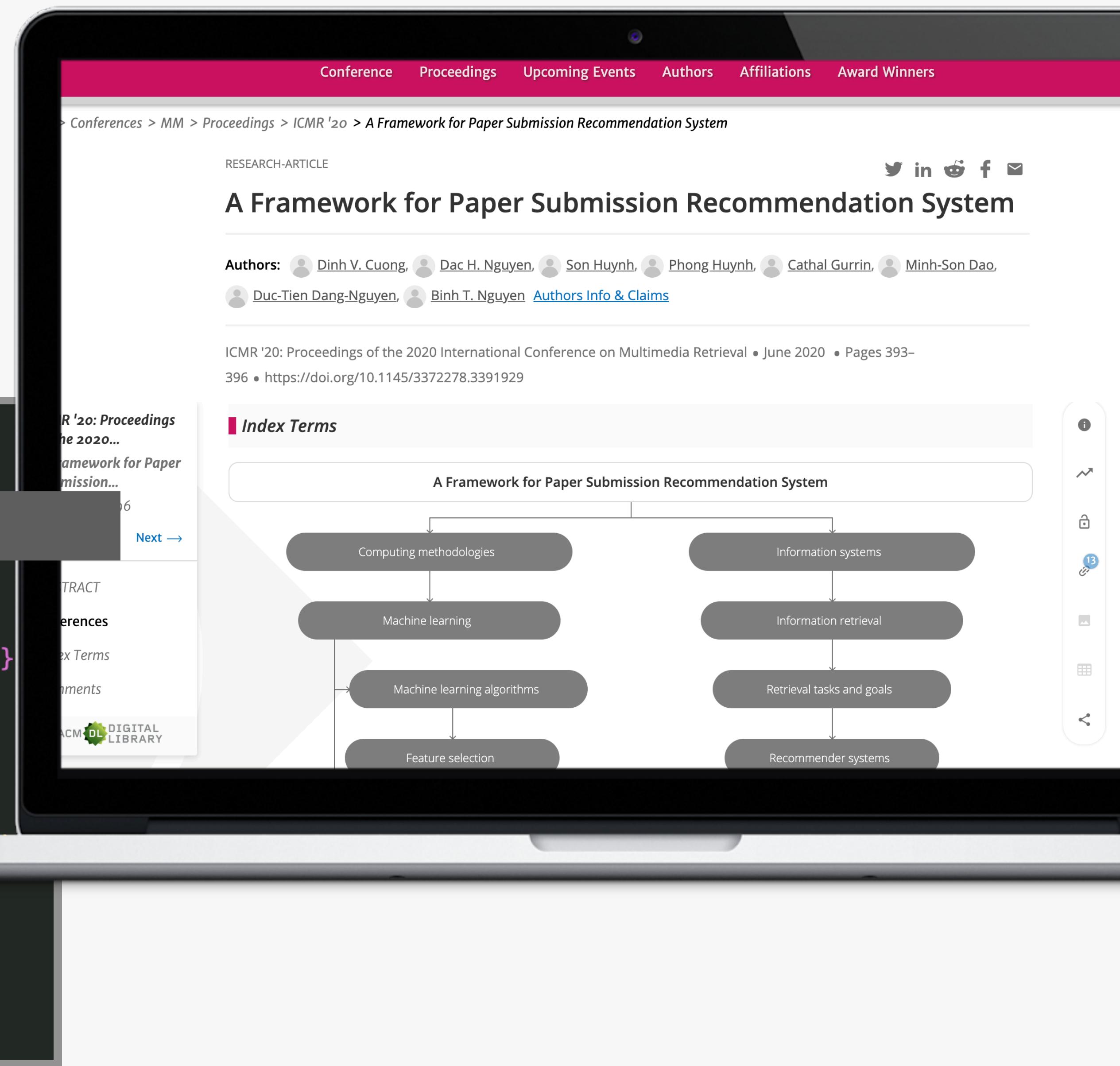
*A lib traz mais informações que a rota direta não traz, mas só pode ser acessada por doi ou paperId

Geração do dataset

Se tem doi pega **keywords** e **autores** do site do doi (vem do próprio dblp)

```
if doi != '':
    resposta = requests.get(doi_url, headers = cabecalho)
    sopa = BeautifulSoup(resposta.text, 'html.parser')
    # print(sopa.prettify())
    author_list = sopa.find('ul',{'aria-label':'authors'})
    if author_list is not None:
        for author in author_list.find_all('a',{'class':'author-name'}):
            authors = authors + author.get('title') + ','
        authors = authors[0:len(authors)-1]
    print(authors)

    kw_list = sopa.find('ol',{'class':'list organizational'})
    if kw_list is not None:
        for kw in kw_list.find_all('a'):
            keywords = keywords + kw.get_text() + ','
        keywords = keywords[0:len(keywords)-1]
    print(keywords)
```



Geração do dataset

Adequa os dados e escreve
em um arquivo **CSV**

```
if article_data is not None:  
    article_data.insert(0,title)  
    article_data.append(year)  
    article_data[3] = authors if authors != '' else article_data[3]  
    article_data[4] = conf  
    if article_data[5] == '' and keywords != '':  
        article_data[5] = keywords  
    elif article_data[5] != '' and keywords != '':  
        article_data[5] = article_data[5] + ', ' + keywords  
    # [title,paperId,doi,authors,publisher,topics,fields_of_study,year]  
    if article_data[6] == '':  
        article_data[6] = genre  
    elif genre.lower() not in article_data[6].lower():  
        article_data[6] = article_data[6] + ', ' + genre  
    else:  
        article_data = [title,'',doi,authors,conf,keywords,genre,year]  
  
writer.writerow(article_data)  
qty_article += 1
```

Dataset > articles-2022-09-19.csv

```
1 title;paperId;doi;authors;publisher;topics;fields_of_study;year  
2 Basil: Breaking up BFT with ACID (transactions).;e3b8493001da3d4bb0c836252d669db25d2b752;10.1145/3471841  
3 Bidl: A High-throughput, Low-latency Permissioned Blockchain Framework for Datacenter Networks.;1841348  
4 Kauri: Scalable BFT Consensus with Pipelined Tree-Based Dissemination and Aggregation.;7f7fe36b216862  
5 iGUARD: In-GPU Advanced Race Detection.;7dfdc9010d42636f50f5462528d0e1e31f0d34c3;10.1145/3477132.3483  
6 Snowboard: Finding Kernel Concurrency Bugs through Systematic Inter-thread Communication Analysis.;59  
7 Rudra: Finding Memory Safety Bugs in Rust at the Ecosystem Scale.;1274ec8320e9ef5e0a2b1db8f665e4a4d1c  
8 Witcher: Systematic Crash Consistency Testing for Non-Volatile Memory Key-Value Stores.;238c4a20cf71f  
9 Understanding and Detecting Software Upgrade Failures in Distributed Systems.;32b6c3fdd1dc10428817503  
10 Crash Consistent Non-Volatile Memory Express.;45704a197ab4600e334d81fe7bc91c992c8f8557;10.1145/347713  
11 Cuckoo Trie: Exploiting Memory-Level Parallelism for Efficient DRAM Indexing.;10cac1ea2fbf8aa65b8404c  
12 Regular Sequential Serializability and Regular Sequential Consistency.;df178ad5801ccfb976bbfd2857e268  
13 Caracal: Contention Management with Deterministic Concurrency Control.;6cad3810d1e6998ec6d713e5c2fdea  
14 The Demikernel Datapath OS Architecture for Microsecond-scale Datacenter Systems.;98899900477479166c0  
15 Birds of a Feather Flock Together: Scaling RDMA RPCs with Flock.;102f6c26581ab7341d72a4022495e007f9f1  
16 RISM: Rethinking the RDMA Interface for Distributed Systems.;4964684f2725abc5a2e1797f57f7fd64da8c439  
17 Kangaroo: Caching Billions of Tiny Objects on Flash.;58b144f3438faa6939309fa4e51dd113dd126c8e;10.1145/3471841  
18 LODA: A Host/Device Co-Design for Strong Predictability Contract on Modern Flash Storage.;98a1764b4c0  
19 FragPicker: A New Defragmentation Tool for Modern Storage Devices.;2feffc076ade1e1d71bc041b82414ef84f  
20 dSpace: Composable Abstractions for Smart Spaces.;03b31dbc09ea53b475c18b827a295196f57677b2;10.1145/3471841  
21 Random Walks on Huge Graphs at Cache Efficiency.;901db9121b32d541e3aedcf315fc85c55a183928;10.1145/3471841  
22 Mycelium: Large-Scale Distributed Graph Queries with Differential Privacy.;2c3f256695ed45668e6d373ab1  
23 HEALER: Relation Learning Guided Kernel Fuzzing.;2c7ff575588c58c85e68a045a5313d61ed520290;10.1145/3471841  
24 Gradient Compression Supercharged High-Performance Data Parallel DNN Training.;1d200ce7d40a24a3e77eee  
25 Generating Complex, Realistic Cloud Workloads using Recurrent Neural Networks.;ad06b571f84e11fc7ff8a3  
26 HeMem: Scalable Tiered Memory Management for Big Data Applications and Real NVM.;eb028f2f10fe0b26a096
```

Geração do dataset

Adequa os dados e escreve
em um arquivo **CSV**

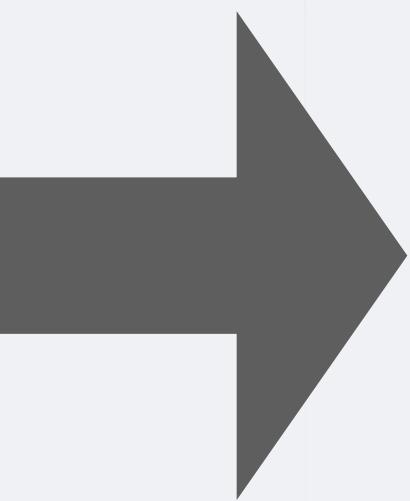
```
> articles-2022-09-19.csv
title;paperId;doi;authors;publisher;topics;fields_of_study;year
Basil: Breaking up BFT with ACID (transactions).;e3b8493001da3d4bbd0c836252d669db25d2b752;10.1145/3471
Bidl: A High-throughput, Low-latency Permissioned Blockchain Framework for Datacenter Networks.;184138
Kauri: Scalable BFT Consensus with Pipelined Tree-Based Dissemination and Aggregation.;7f7fe36b2168628
iGUARD: In-GPU Advanced Race Detection.;7dfdc9010d42636f50f5462528d0e1e31f0d34c3;10.1145/3477132.3483
Snowboard: Finding Kernel Concurrency Bugs through Systematic Inter-thread Communication Analysis.;592
Rudra: Finding Memory Safety Bugs in Rust at the Ecosystem Scale.;1274ec8320e9ef5e0a2b1db8f665e4a4d1d
Witcher: Systematic Crash Consistency Testing for Non-Volatile Memory Key-Value Stores.;238c4a20cf71f
Understanding and Detecting Software Upgrade Failures in Distributed Systems.;32b6c3fdd1dc10428817503
Crash Consistent Non-Volatile Memory Express.;45704a197ab4600e334d81fe7bc91c992c8f8557;10.1145/3477132
Cuckoo Trie: Exploiting Memory-Level Parallelism for Efficient DRAM Indexing.;10cac1ea2fbf8aa65b8404d0
Regular Sequential Serializability and Regular Sequential Consistency.;df178ad5801ccfb976bbfd2857e268a
Caracal: Contention Management with Deterministic Concurrency Control.;6cad3810d1e6998ec6d713e5c2fdea
The Demikernel Datapath OS Architecture for Microsecond-scale Datacenter Systems.;98899900477479166c0
Birds of a Feather Flock Together: Scaling RDMA RPCs with Flock.;102f6c26581ab7341d72a4022495e007f9f12
PRISM: Rethinking the RDMA Interface for Distributed Systems.;4964684f2725abc5a2e1797f57f7fd64da8c439
Kangaroo: Caching Billions of Tiny Objects on Flash.;58b144f3438faa6939309fa4e51dd113dd126c8e;10.1145/3477132
LODA: A Host/Device Co-Design for Strong Predictability Contract on Modern Flash Storage.;98a1764b4c0a
FragPicker: A New Defragmentation Tool for Modern Storage Devices.;2fefffc076ade1e1d71bc041b82414ef84f
dSpace: Composable Abstractions for Smart Spaces.;03b31dbc09ea53b475c18b827a295196f57677b2;10.1145/3477132
Random Walks on Huge Graphs at Cache Efficiency.;901db9121b32d541e3aedcf315fc85c55a183928;10.1145/3477132
Mycelium: Large-Scale Distributed Graph Queries with Differential Privacy.;2c3f256695ed45668e6d373ab1
HEALER: Relation Learning Guided Kernel Fuzzing.;2c7ff575588c58c85e68a045a5313d61ed520290;10.1145/3477132
Gradient Compression Supercharged High-Performance Data Parallel DNN Training.;1d200ce7d40a24a3e77eee
Generating Complex, Realistic Cloud Workloads using Recurrent Neural Networks.;ad06b571f84e11fc7ff8a3
HeMem: Scalable Tiered Memory Management for Big Data Applications and Real NVM.;eb028f2f10fe0b26a096
```



Geração do dataset

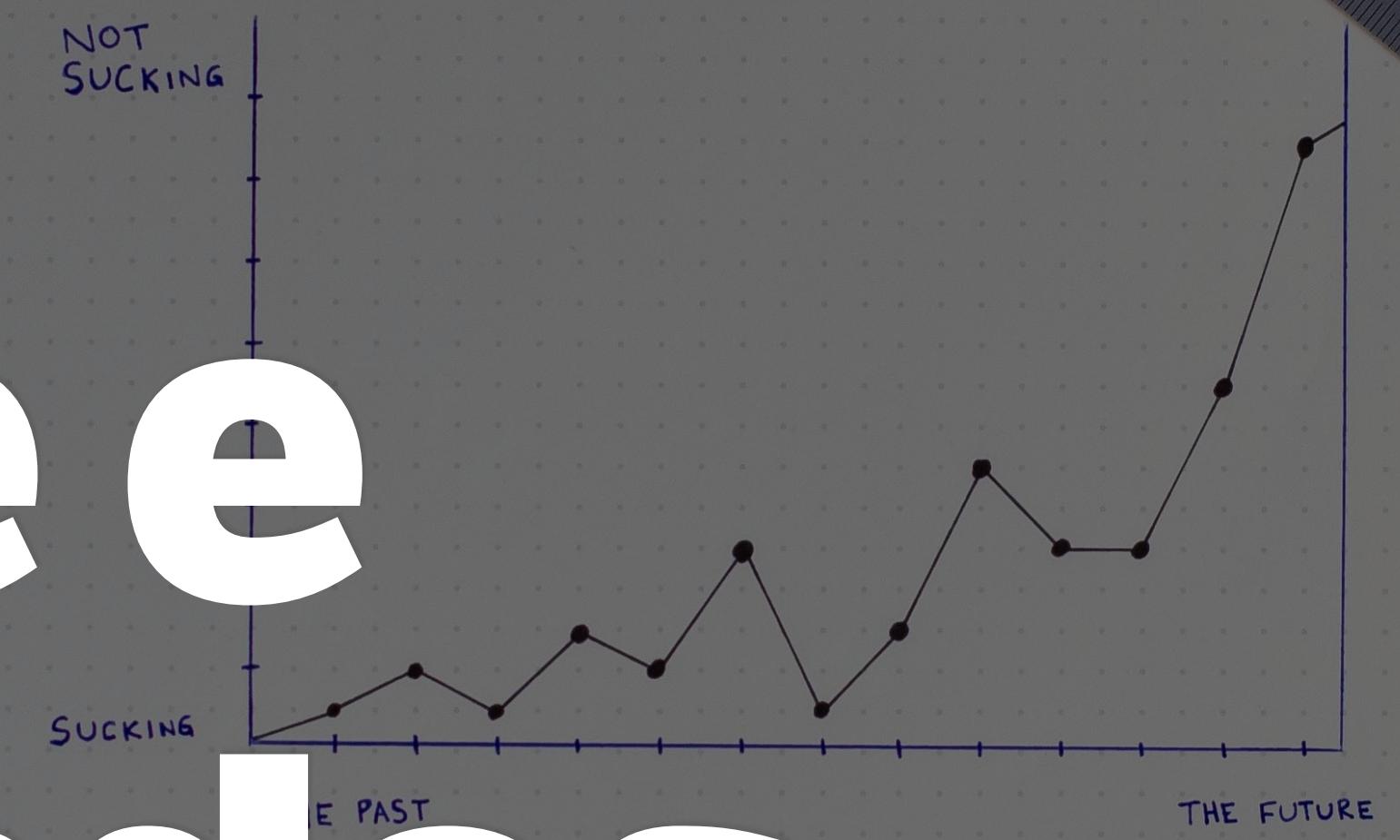
Adequa os dados e escreve
em um arquivo **CSV**

```
> articles-2022-09-19.csv
title;paperId;doi;authors;publisher;topics;fields_of_study;year
Basil: Breaking up BFT with ACID (transactions).;e3b8493001da3d4bbd0c836252d669db25d2b752;10.1145/3471
Bidl: A High-throughput, Low-latency Permissioned Blockchain Framework for Datacenter Networks.;184138
Kauri: Scalable BFT Consensus with Pipelined Tree-Based Dissemination and Aggregation.;7f7fe36b2168628
iGUARD: In-GPU Advanced Race Detection.;7dfdc9010d42636f50f5462528d0e1e31f0d34c3;10.1145/3477132.3483
Snowboard: Finding Kernel Concurrency Bugs through Systematic Inter-thread Communication Analysis.;592
Rudra: Finding Memory Safety Bugs in Rust at the Ecosystem Scale.;1274ec8320e9ef5e0a2b1db8f665e4a4d1d
Witcher: Systematic Crash Consistency Testing for Non-Volatile Memory Key-Value Stores.;238c4a20cf71f
Understanding and Detecting Software Upgrade Failures in Distributed Systems.;32b6c3fdd1dc10428817503
Crash Consistent Non-Volatile Memory Express.;45704a197ab4600e334d81fe7bc91c992c8f8557;10.1145/3477132
Cuckoo Trie: Exploiting Memory-Level Parallelism for Efficient DRAM Indexing.;10cac1ea2fbf8aa65b8404d0
Regular Sequential Serializability and Regular Sequential Consistency.;df178ad5801ccfb976bbfd2857e268a
Caracal: Contention Management with Deterministic Concurrency Control.;6cad3810d1e6998ec6d713e5c2fdea
The Demikernel Datapath OS Architecture for Microsecond-scale Datacenter Systems.;98899900477479166c0
Birds of a Feather Flock Together: Scaling RDMA RPCs with Flock.;102f6c26581ab7341d72a4022495e007f9f12
PRISM: Rethinking the RDMA Interface for Distributed Systems.;4964684f2725abc5a2e1797f57f7fd64da8c439
Kangaroo: Caching Billions of Tiny Objects on Flash.;58b144f3438faa6939309fa4e51dd113dd126c8e;10.1145
lODA: A Host/Device Co-Design for Strong Predictability Contract on Modern Flash Storage.;98a1764b4c0a
FragPicker: A New Defragmentation Tool for Modern Storage Devices.;2fefffc076ade1e1d71bc041b82414ef84f
dSpace: Composable Abstractions for Smart Spaces.;03b31dbc09ea53b475c18b827a295196f57677b2;10.1145/347
Random Walks on Huge Graphs at Cache Efficiency.;901db9121b32d541e3aedcf315fc85c55a183928;10.1145/347
Mycelium: Large-Scale Distributed Graph Queries with Differential Privacy.;2c3f256695ed45668e6d373ab1
HEALER: Relation Learning Guided Kernel Fuzzing.;2c7ff575588c58c85e68a045a5313d61ed520290;10.1145/347
Gradient Compression Supercharged High-Performance Data Parallel DNN Training.;1d200ce7d40a24a3e77eee
Generating Complex, Realistic Cloud Workloads using Recurrent Neural Networks.;ad06b571f84e11fc7ff8a3
HeMem: Scalable Tiered Memory Management for Big Data Applications and Real NVM.;eb028f2f10fe0b26a096
```



2783 artigos
3597 tópicos

Análise e resultados



Análise e Resultados



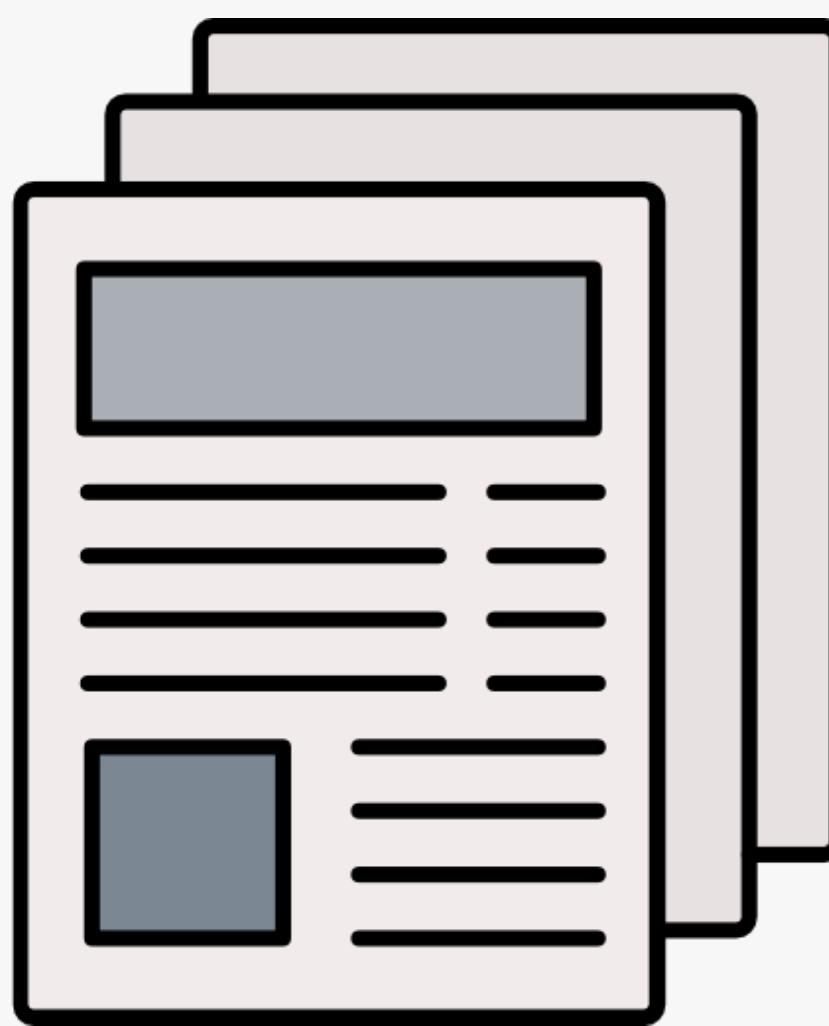
2783 artigos

10 conferências

Entre 2018 e 2022

3597 tópicos

Análise e Resultados



Quantidade de tópicos por conferência:

Tabela 2. Quantidade de tópicos por conferência

Conferência	Quantidade de Tópicos
SOSP	383
OSDI	323
NDSS	1253
Mobihoc	628
SIGCOMM	485
SenSys	444
MOBICOM	702
CIDR	421
USENIX Security Symposium	1343
EUROCRYPT	395

Análise e Resultados



Análises e resultados



Quais são os 6 tópicos mais publicados nos eventos?

Dicionário com a quantidade de vezes que um tópico aparece

```
1 {  
2   ' COMPUTER SYSTEMS ORGANIZATION' : 29,  
3   ' DEPENDABLE AND FAULT-TOLERANT SYSTEMS AND NETWORKS' : 18,  
4   ' SECURITY AND PRIVACY' : 25,  
5   ...  
6 }
```

Análises e resultados



Quais são os 6 tópicos mais publicados nos eventos?

Dicionário com a quantidade de vezes que um tópico aparece

Ordenado de forma decrescente

Pegou os 6 primeiros

Análises e resultados



Quais são os 6 tópicos mais publicados nos eventos?

Tabela 3. Quantidade de artigos de cada um dos 6 maiores tópicos

Tópico	Quantidade de artigos
'EXPERIMENT'	154
'ALGORITHM'	150
'COMPUTER SCIENCE'	131
'SENSOR'	114
'PROTOTYPE'	113
'MACHINE LEARNING'	107

Análises e resultados



Quais são os 6 tópicos mais publicados por ano (de 2018 a 2022) nestes eventos?

Dicionário de anos com a quantidade de vezes que um tópico foi publicado naquele ano

```
1 {  
2   2022: {  
3     'NETWORK CONGESTION': 1,  
4     'ELASTICITY (DATA STORE)': 1,  
5     'THROUGHPUT': 1,  
6     'FREQUENCY RESPONSE': 1,  
7     ...  
8   }  
9 }
```

Análises e resultados



Quais são os 6 tópicos mais publicados por ano (de 2018 a 2022) nestes eventos?

**Dicionário de anos com a quantidade de vezes
que um tópico foi publicado naquele ano**

Ordenado de forma decrescente

Pegou os 6 primeiros de cada ano

Análises e resultados



Quais são os 6 tópicos mais publicados por ano (de 2018 a 2022) nestes eventos?

Tabela 4. Top 6 Tópicos Publicados Em Cada Ano

Ano	Top 6 Tópicos
'2018'	'EXPERIMENT', 'SENSOR', 'PROTOTYPE', 'ALGORITHM', 'OVERHEAD (COMPUTING)', 'THROUGHPUT'
'2019'	'ALGORITHM', 'EXPERIMENT', 'PROTOTYPE', 'MACHINE LEARNING', 'SENSOR', 'SCALABILITY'
'2020'	'COMPUTER SCIENCE', 'ALGORITHM', 'EXPERIMENT', 'NETWORKS', 'MACHINE LEARNING', 'ADVERSARY (CRYPTOGRAPHY)'
'2021'	'COMPUTER SCIENCE', 'NETWORKS', 'SOFTWARE AND ITS ENGINEERING', 'COMPUTER SYSTEMS ORGANIZATION', 'SOFTWARE ORGANIZATION AND PROPERTIES', 'INFORMATION SYSTEMS'
'2022'	'COMPUTER SCIENCE', 'NETWORK CONGESTION', 'ELASTICITY (DATA STORE)', 'THROUGHPUT', 'FREQUENCY RESPONSE', 'FLOW'

Análises e resultados



Para cada tópico, quais as conferências que mais os publicam?

```
1 {  
2   'NP-HARDNESS': {  
3     'MobiHoc': {  
4       2020: 2,  
5       2019: 2,  
6       2018: 2  
7     },  
8     'MOBICOM': {  
9       2018: 1  
10    }  
11  },  
12  'ONLINE LEARNING SETTINGS': {  
13    'MobiHoc': {  
14      2021: 1  
15    }  
16  }  
17  ...  
18 }
```

```
1 {  
2   'TROJAN HORSE (COMPUTING)': {  
3     2018: {  
4       'NDSS': 1,  
5       'USENIX Security Symposium': 1  
6     },  
7     2021: {  
8       'USENIX Security Symposium': 1  
9     },  
10    2020: {  
11      'USENIX Security Symposium': 1  
12    },  
13    2019: {  
14      'USENIX Security Symposium': 3  
15    }  
16  },  
17  ...
```

Análises e resultados



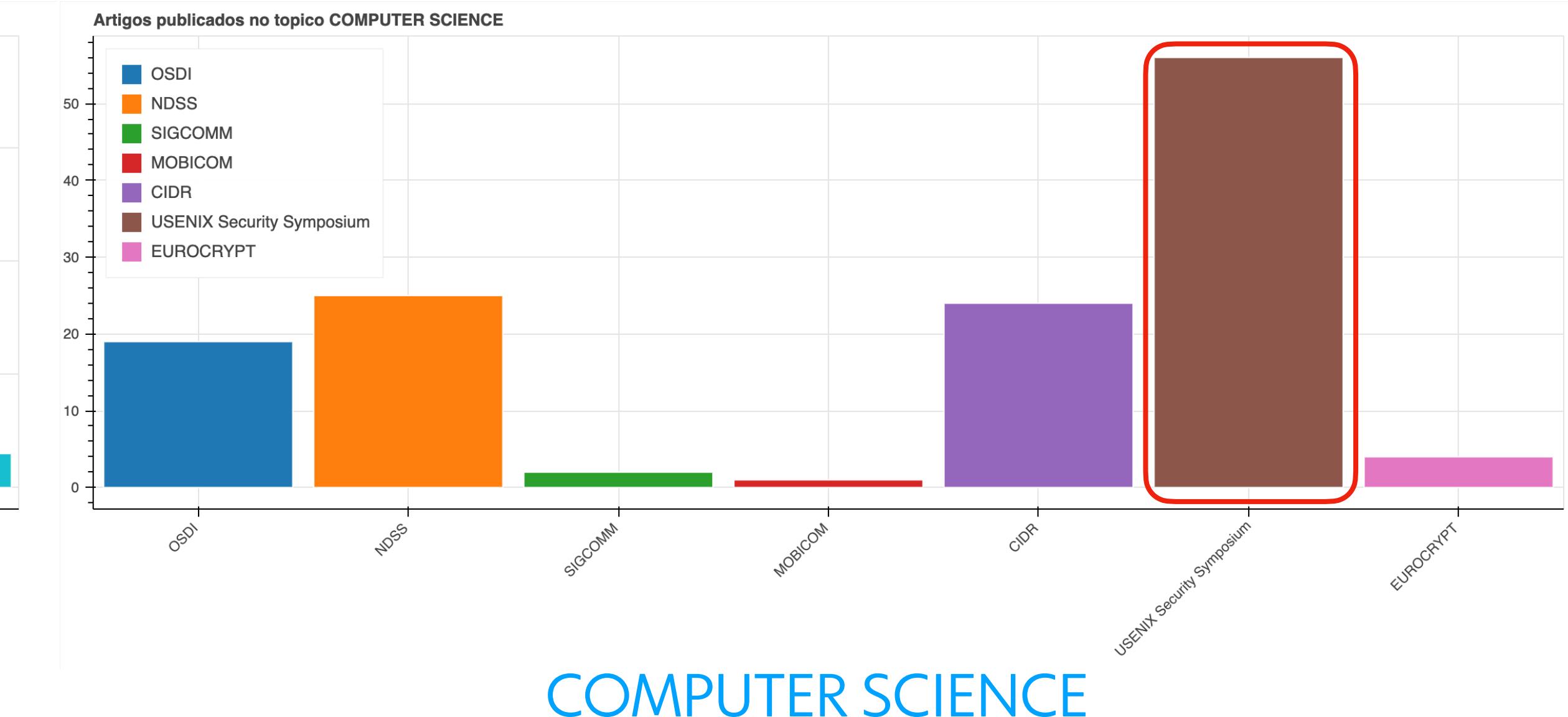
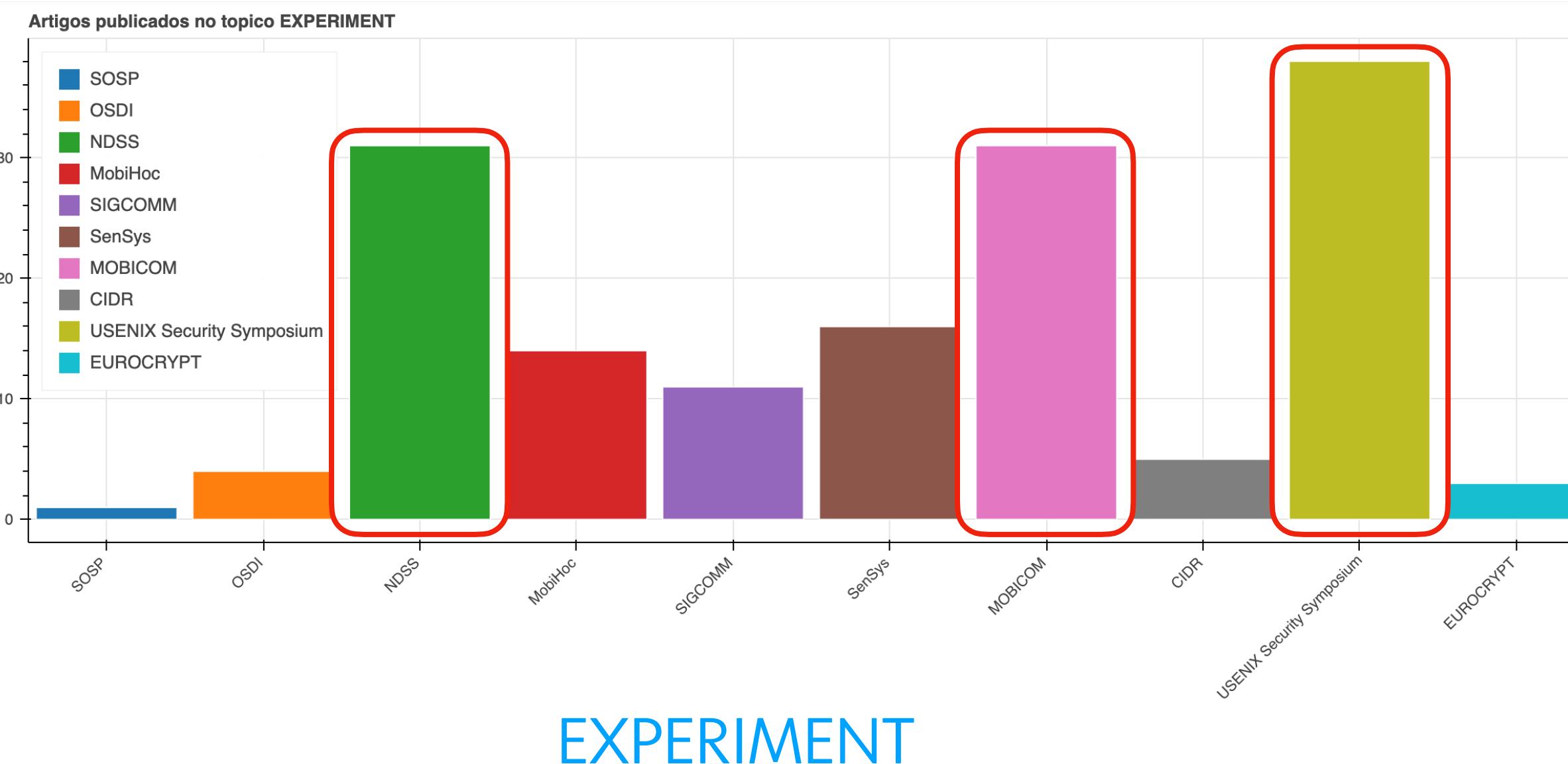
Para cada tópico, quais as conferências que mais os publicam?

Para cada tópico, faz a soma da quantidade de publicações por conferência

Análises e resultados



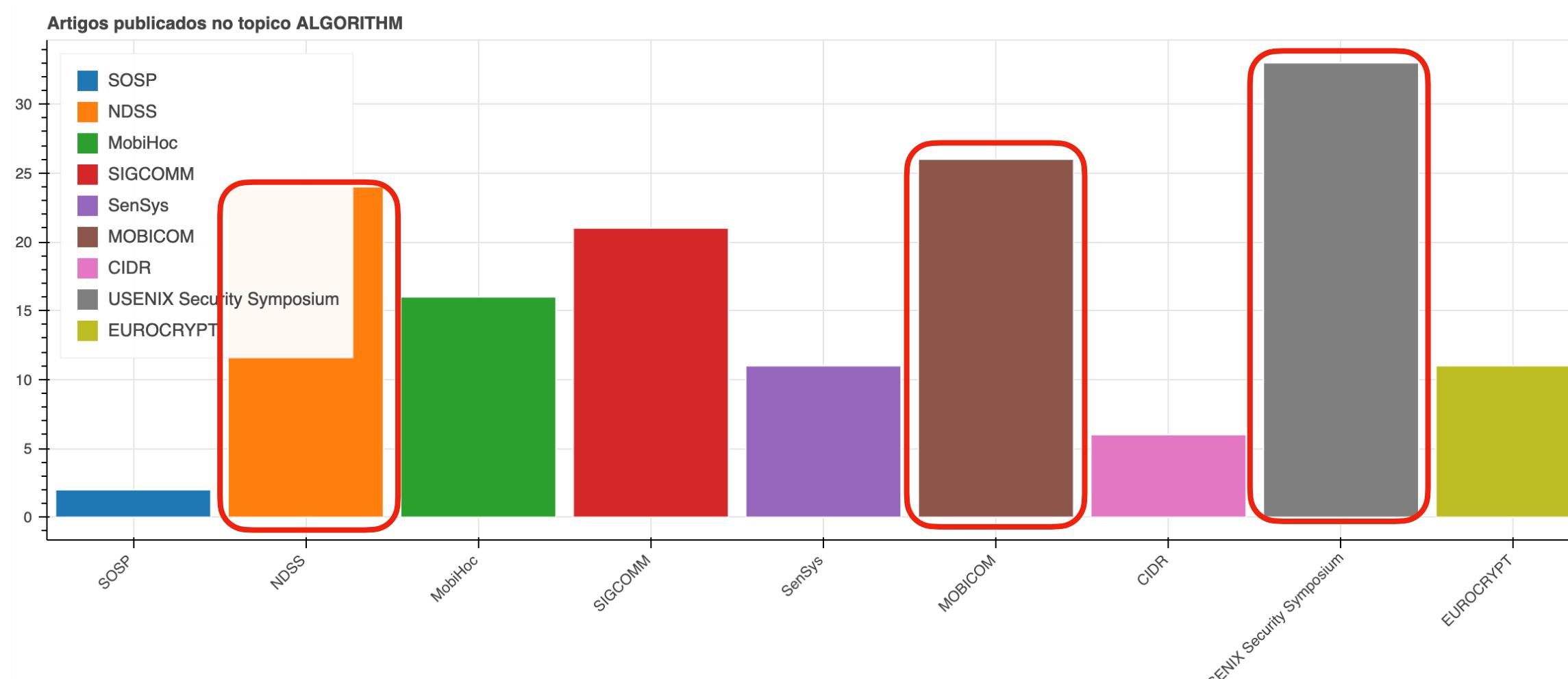
Para cada tópico, quais as conferências que mais os publicam?



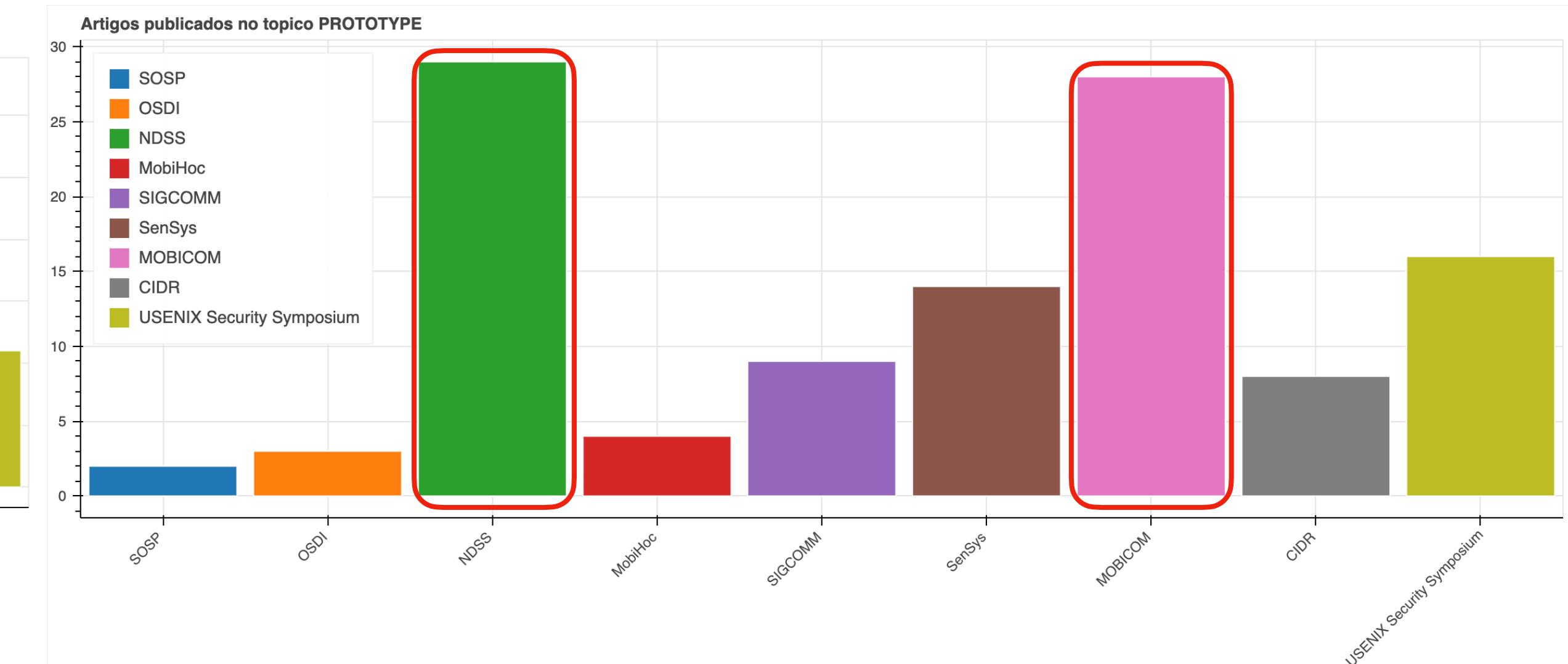
Análises e resultados



Para cada tópico, quais as conferências que mais os publicam?



ALGORITHM

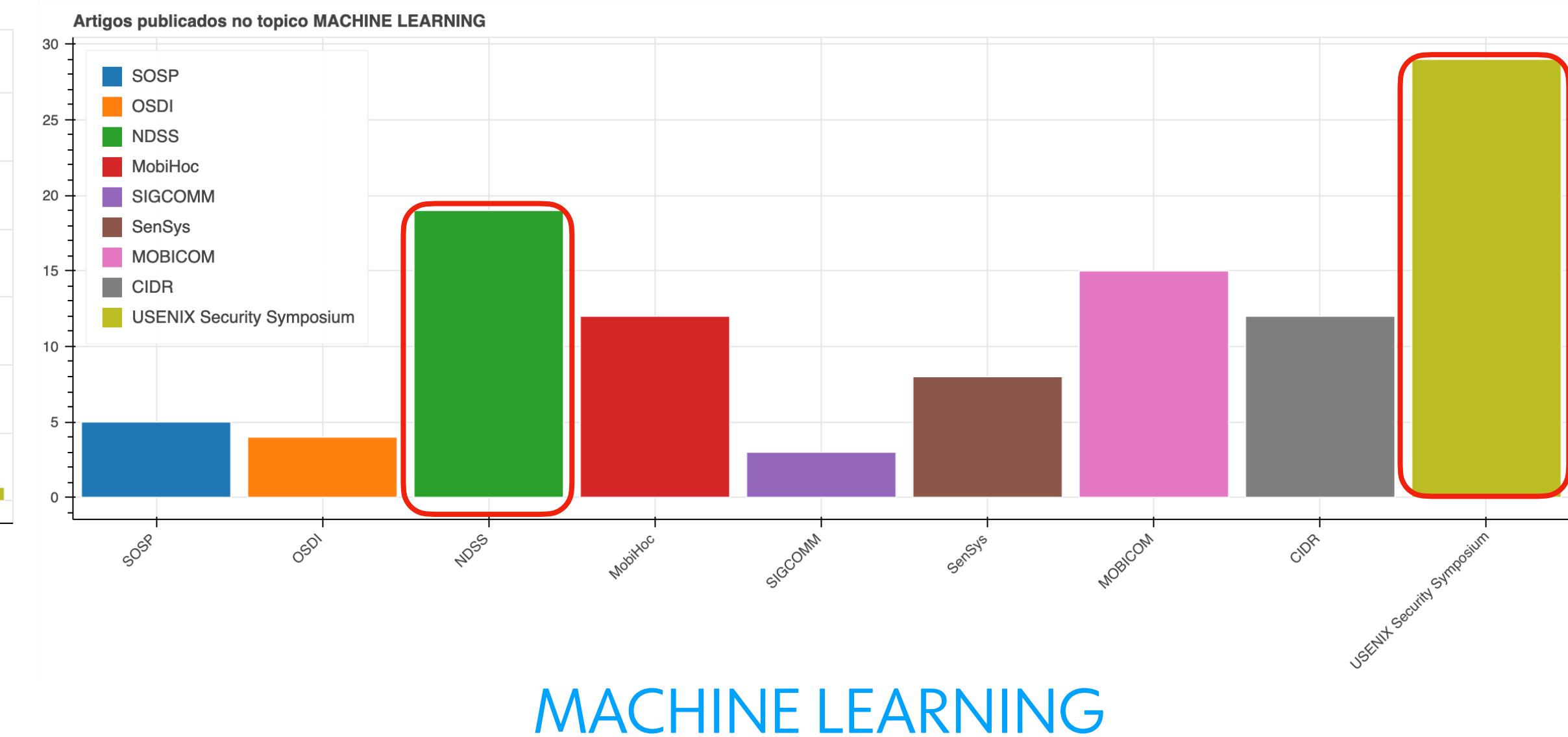
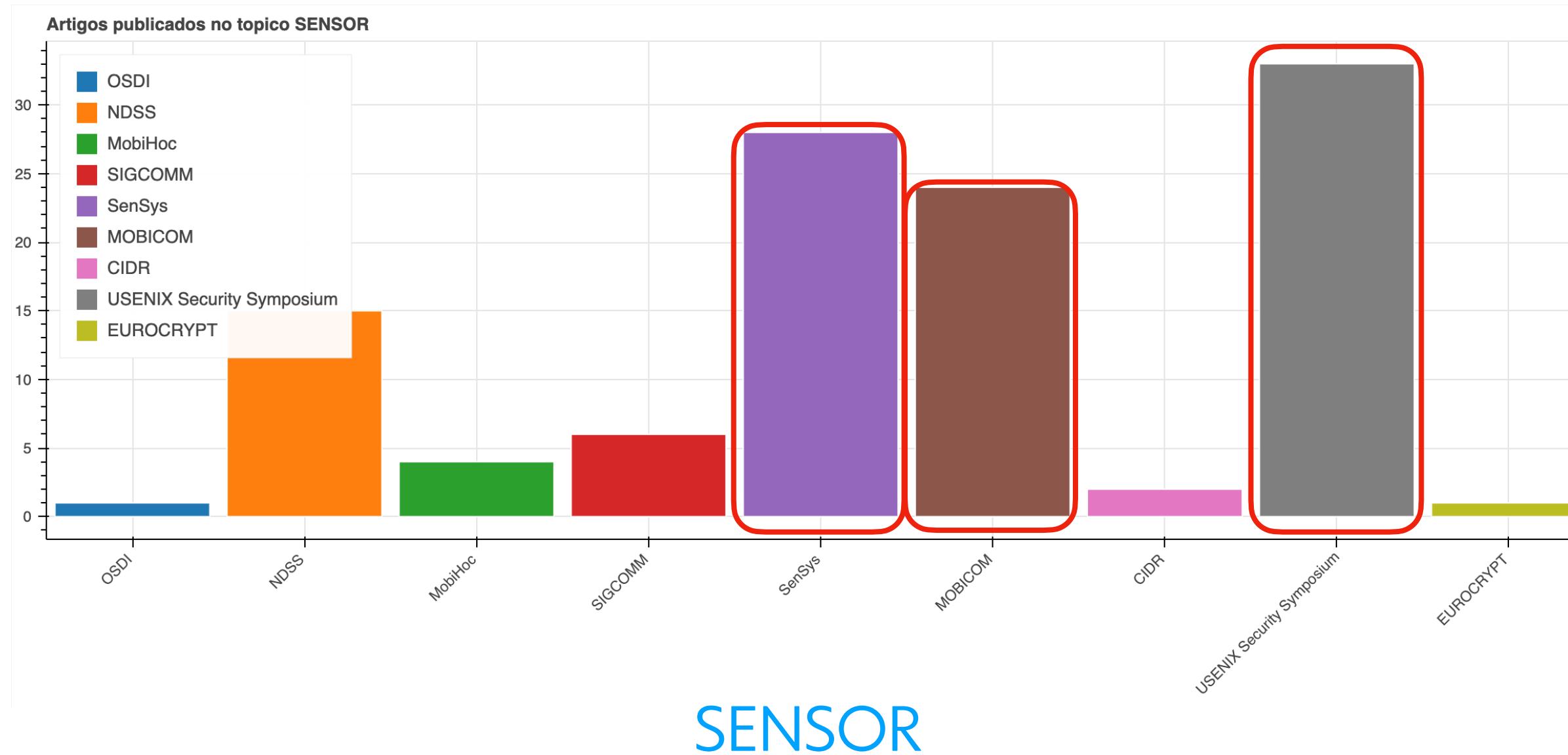


PROTOTYPE

Análises e resultados



Para cada tópico, quais as conferências que mais os publicam?



Análises e resultados



Para cada tópico, quais as conferências que mais os publicam?

Publicações do último ano: Ordenar o ano de forma crescente e pegar a última linha para todas as conferências (de cada tópico)

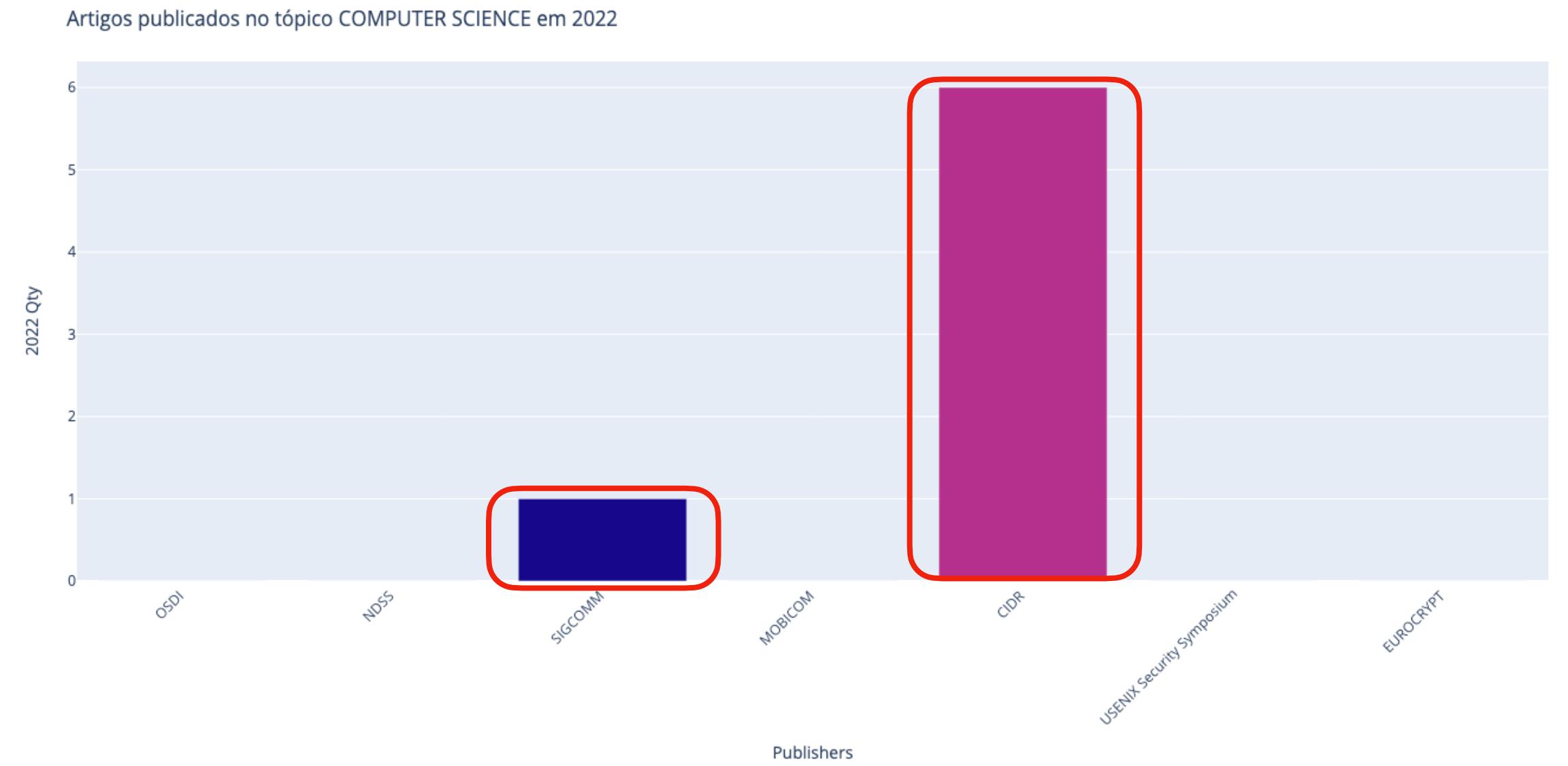
Análises e resultados



Para cada tópico, quais as conferências que mais os publicam?



EXPERIMENT

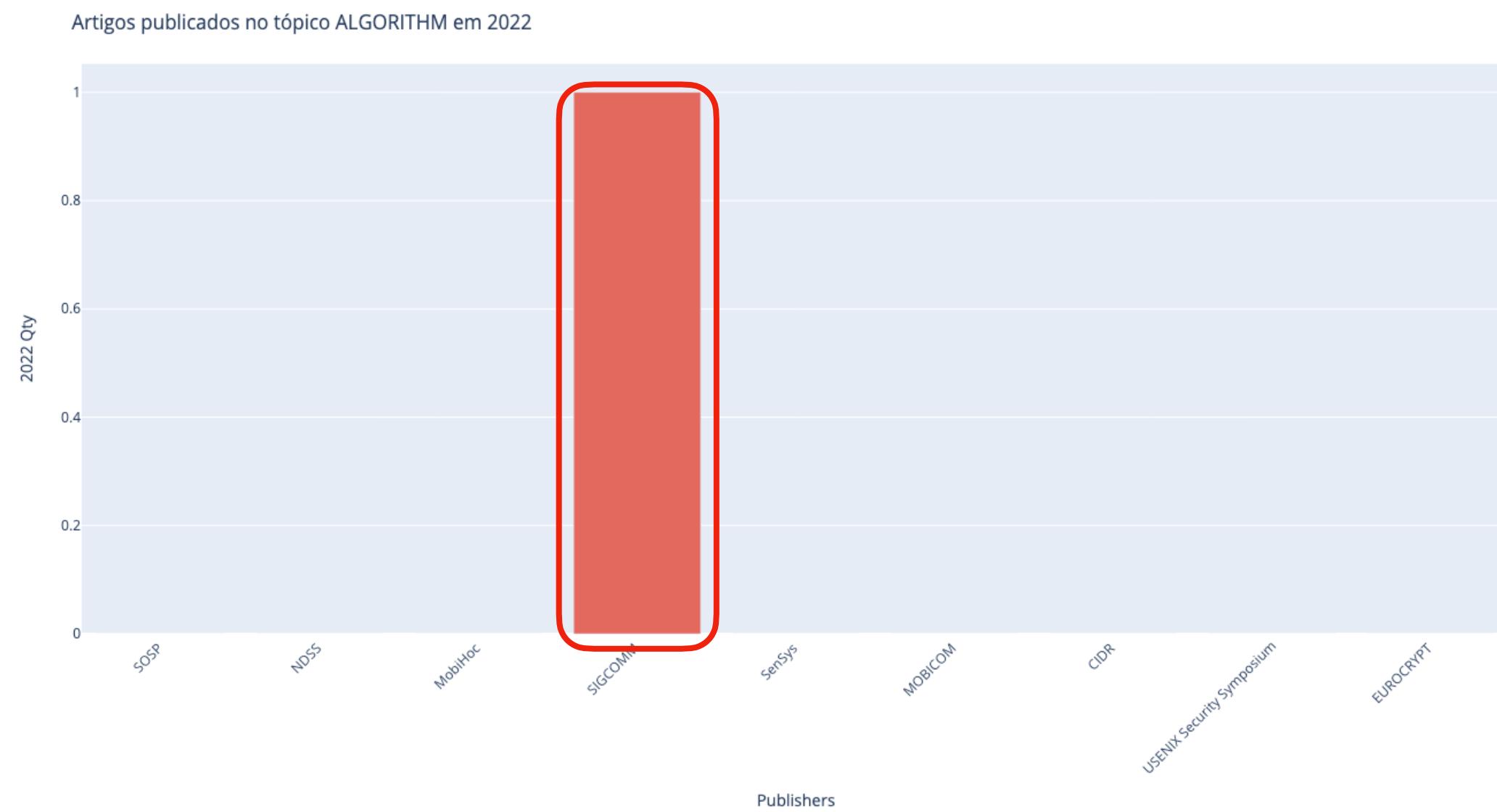


COMPUTER SCIENCE

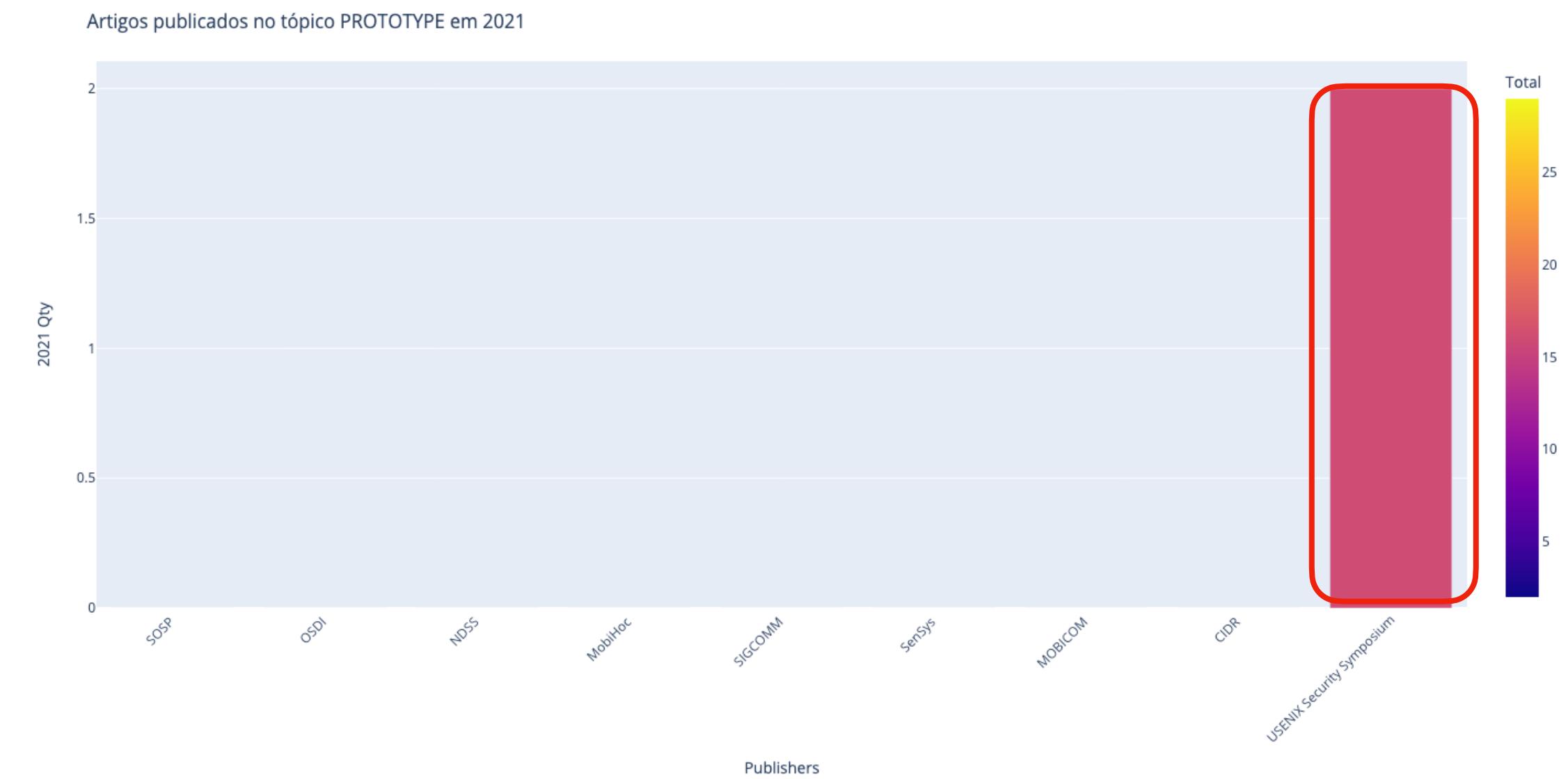
Análises e resultados



Para cada tópico, quais as conferências que mais os publicam?



ALGORITHM



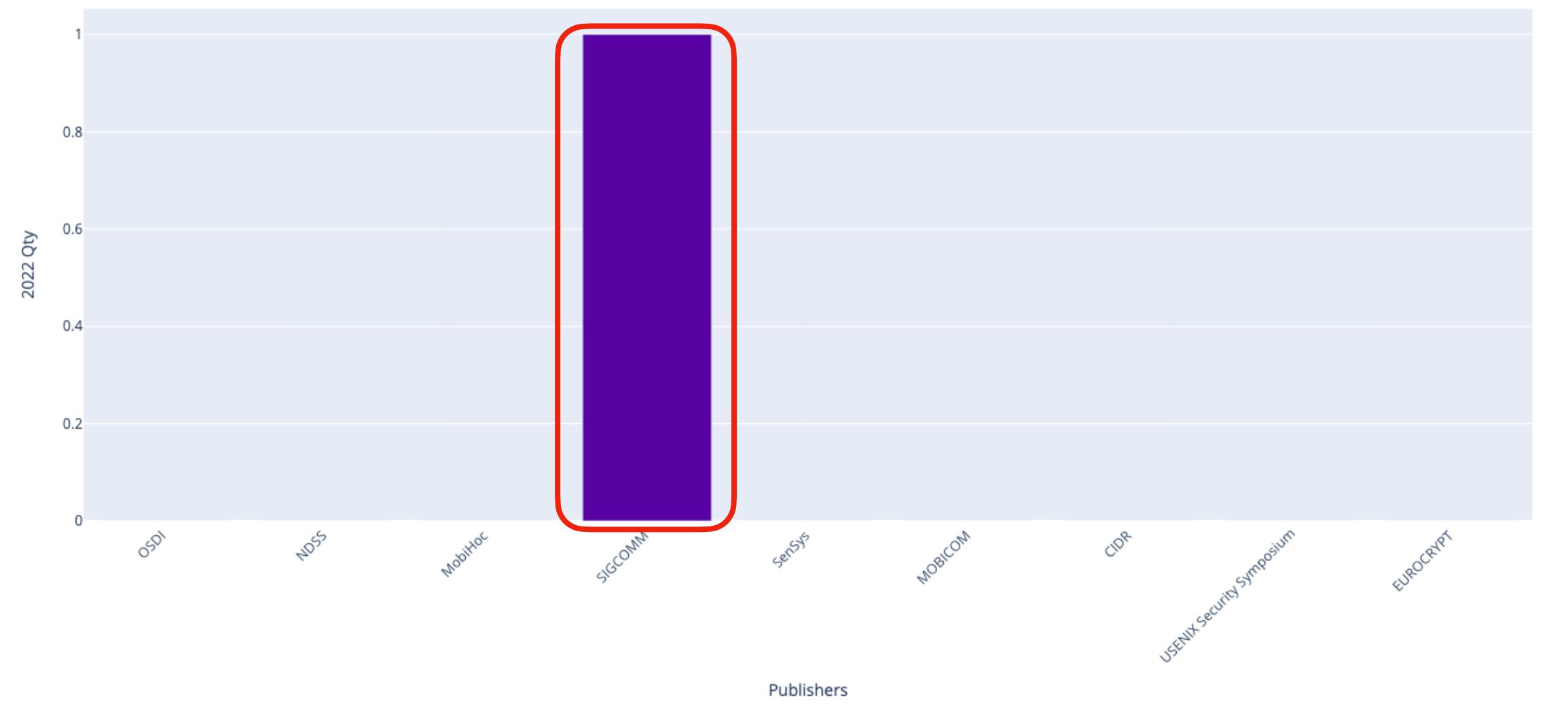
PROTOTYPE

Análises e resultados



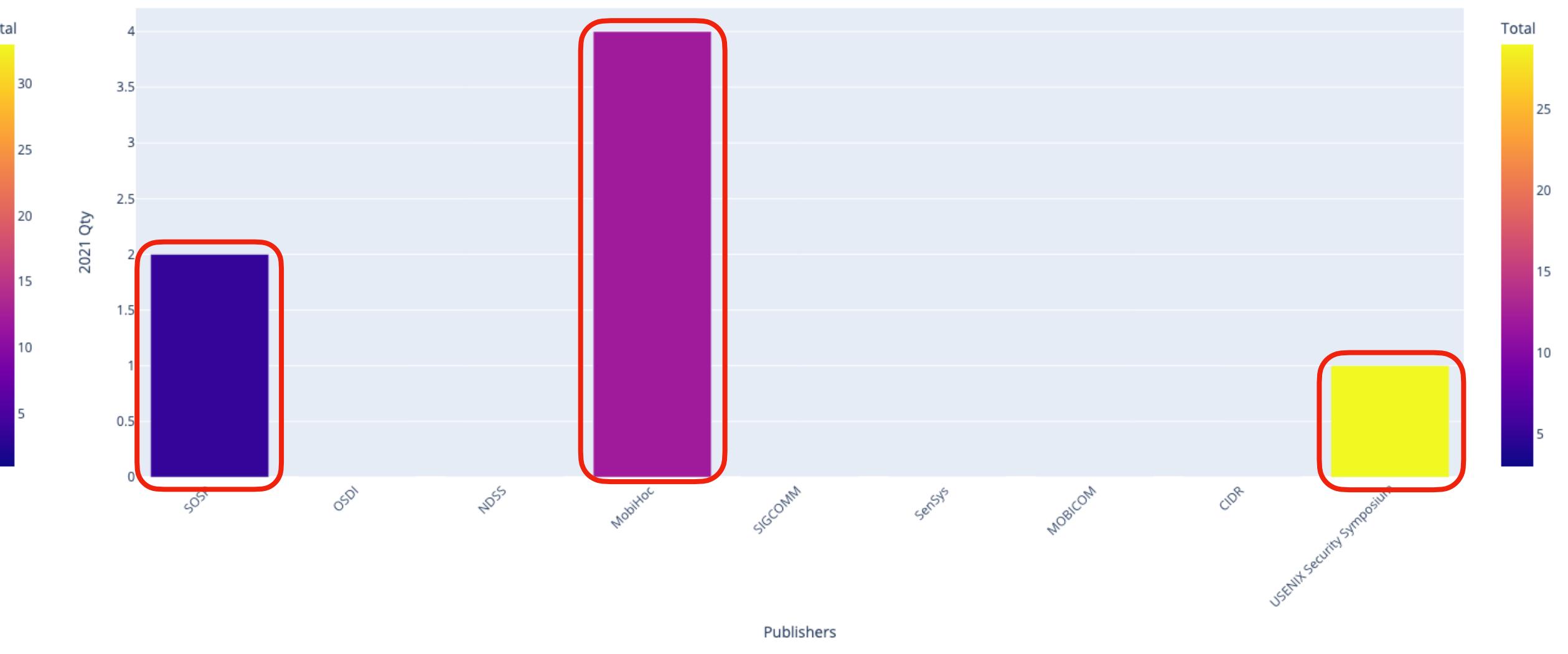
Para cada tópico, quais as conferências que mais os publicam?

Artigos publicados no tópico SENSOR em 2022



SENSOR

Artigos publicados no tópico MACHINE LEARNING em 2021



MACHINE LEARNING

Análises e resultados



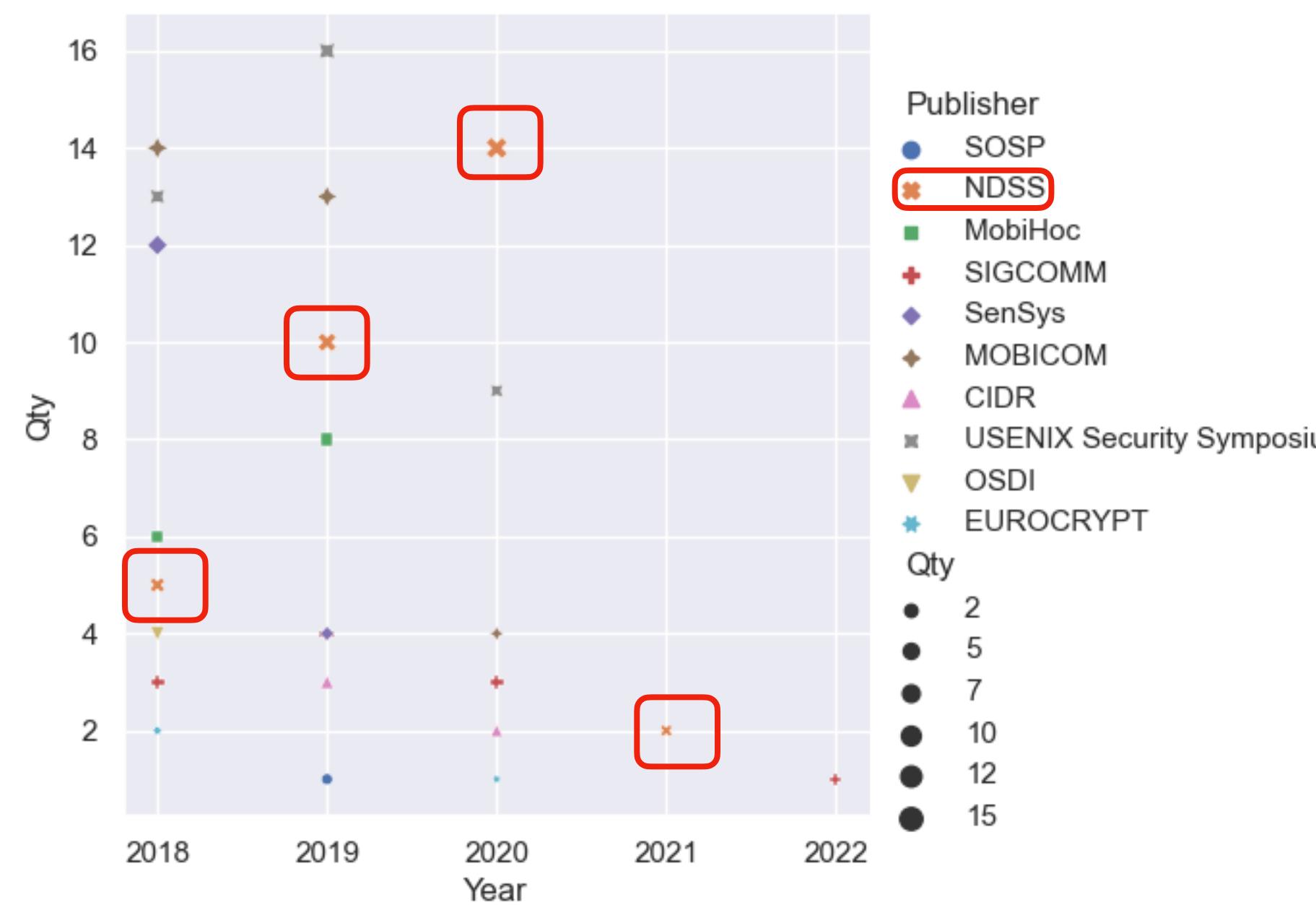
Como foram as evoluções para cada tópico das publicações das conferências ao longo dos anos?

Para cada tópico pega o quanto cada conferência publicou em cada ano

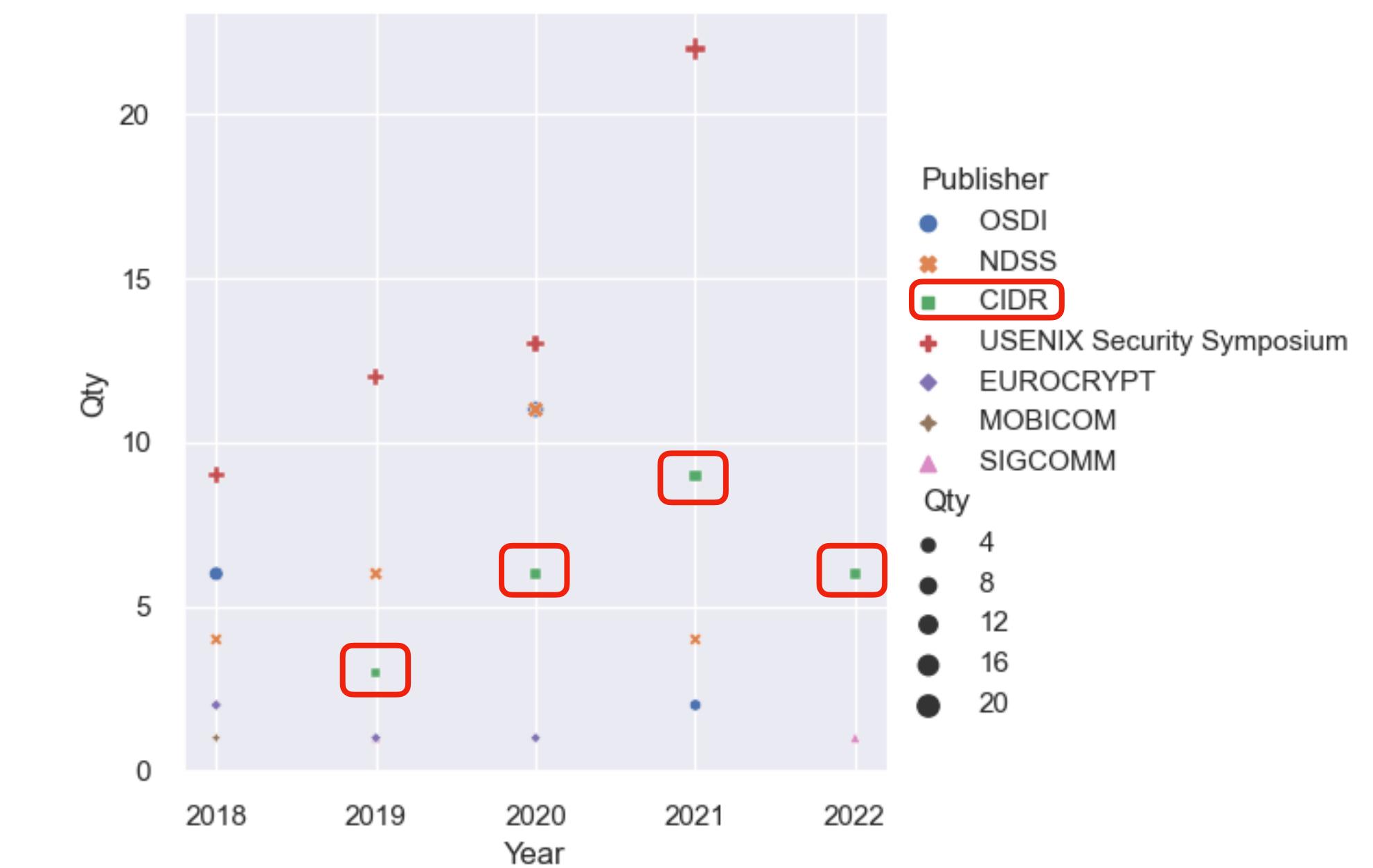
Análises e resultados



Como foram as evoluções ao longo dos anos de publicações das conferências para cada tópico?



EXPERIMENT

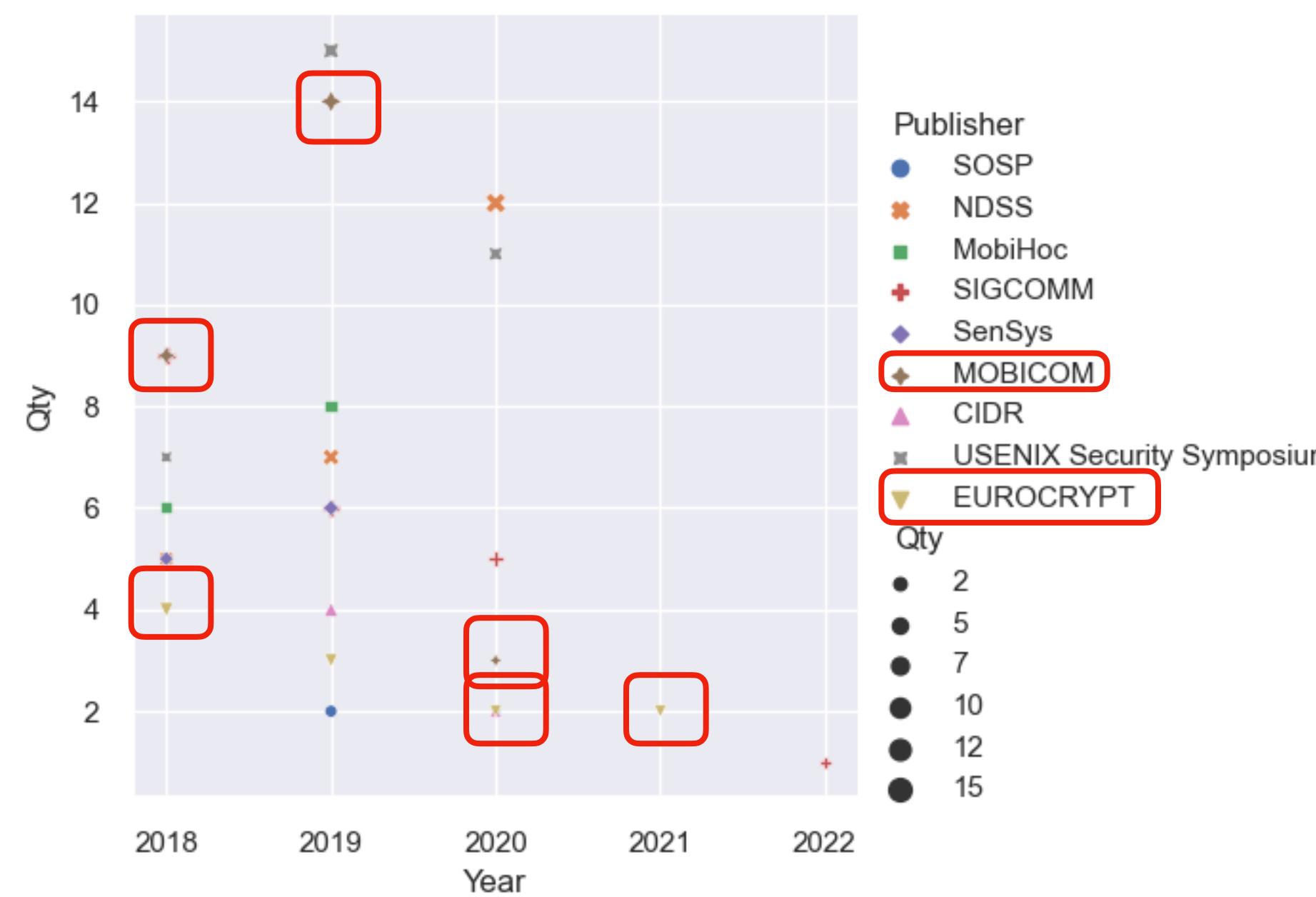


COMPUTER SCIENCE

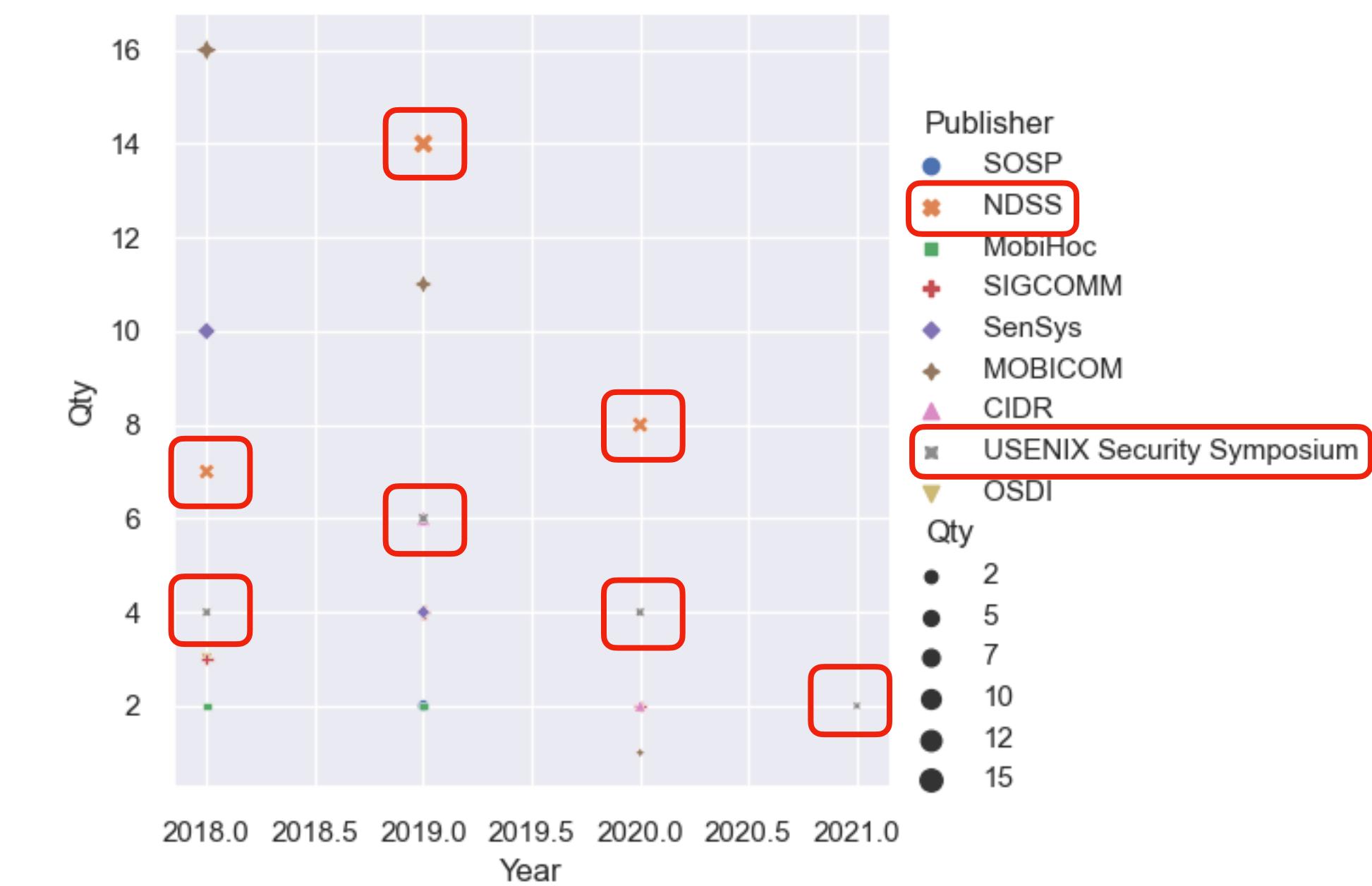
Análises e resultados



Como foram as evoluções ao longo dos anos de publicações das conferências para cada tópico?



ALGORITHM

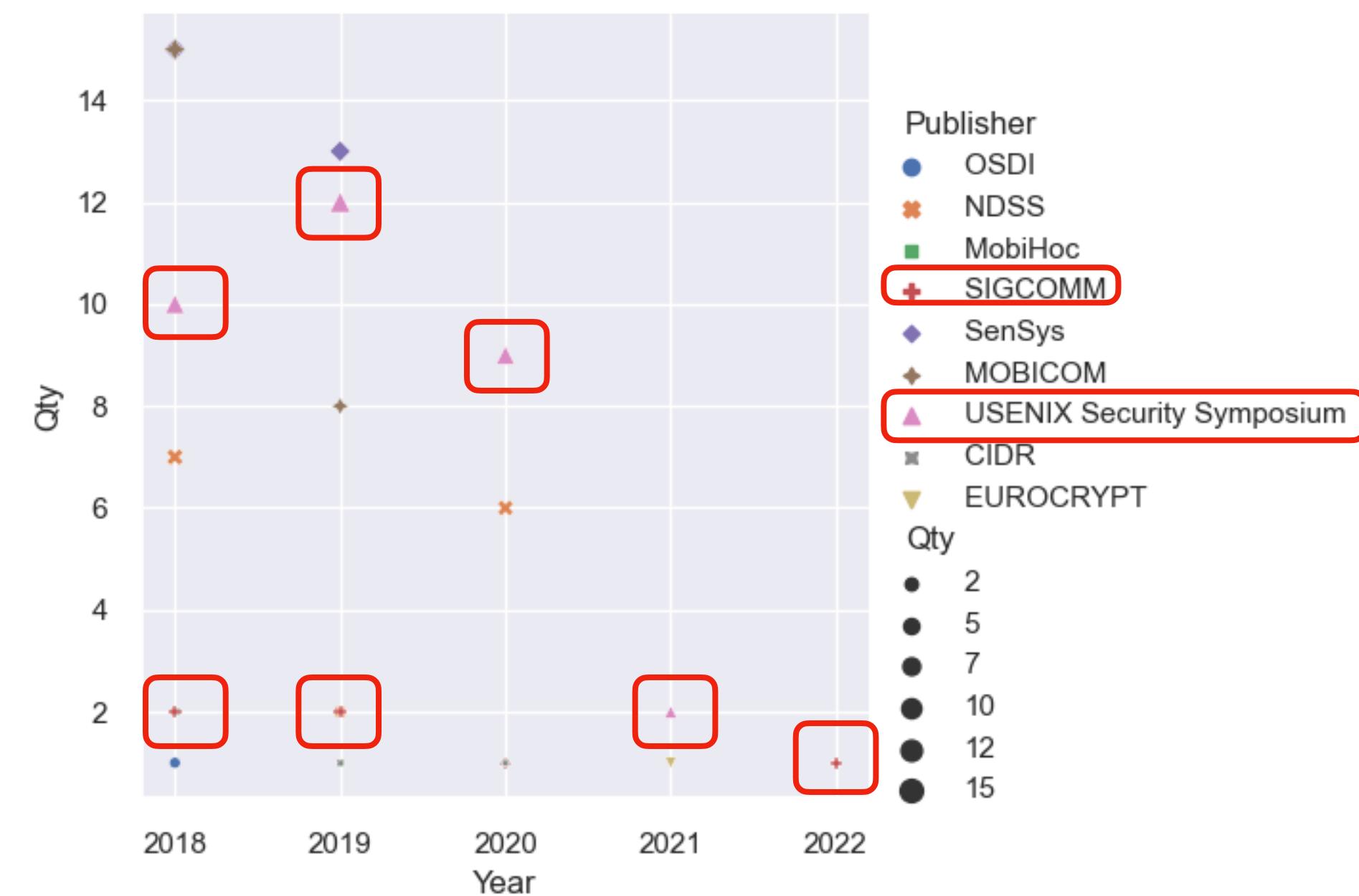


PROTOTYPE

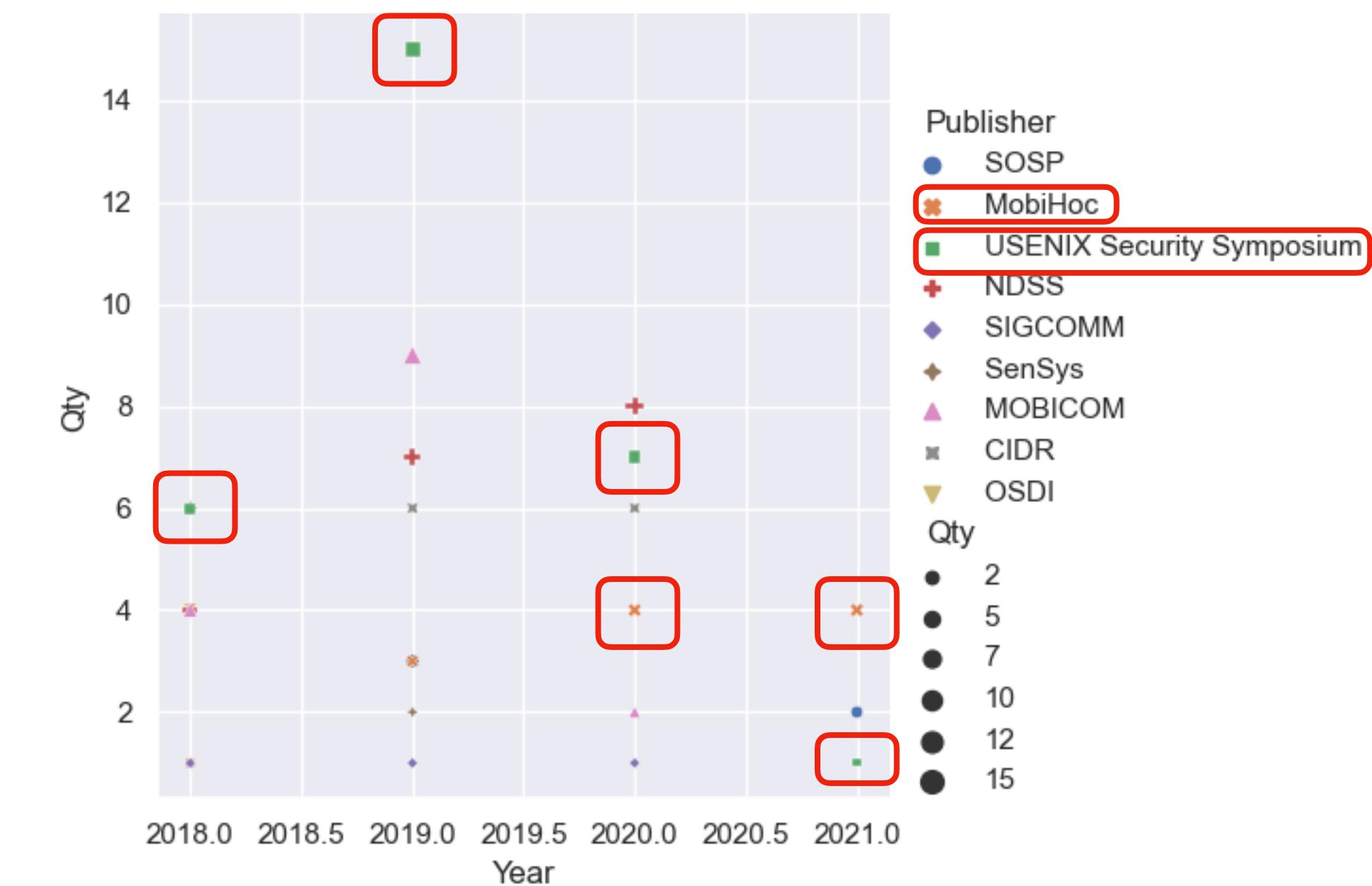
Análises e resultados



Como foram as evoluções ao longo dos anos de publicações das conferências para cada tópico?



SENSOR

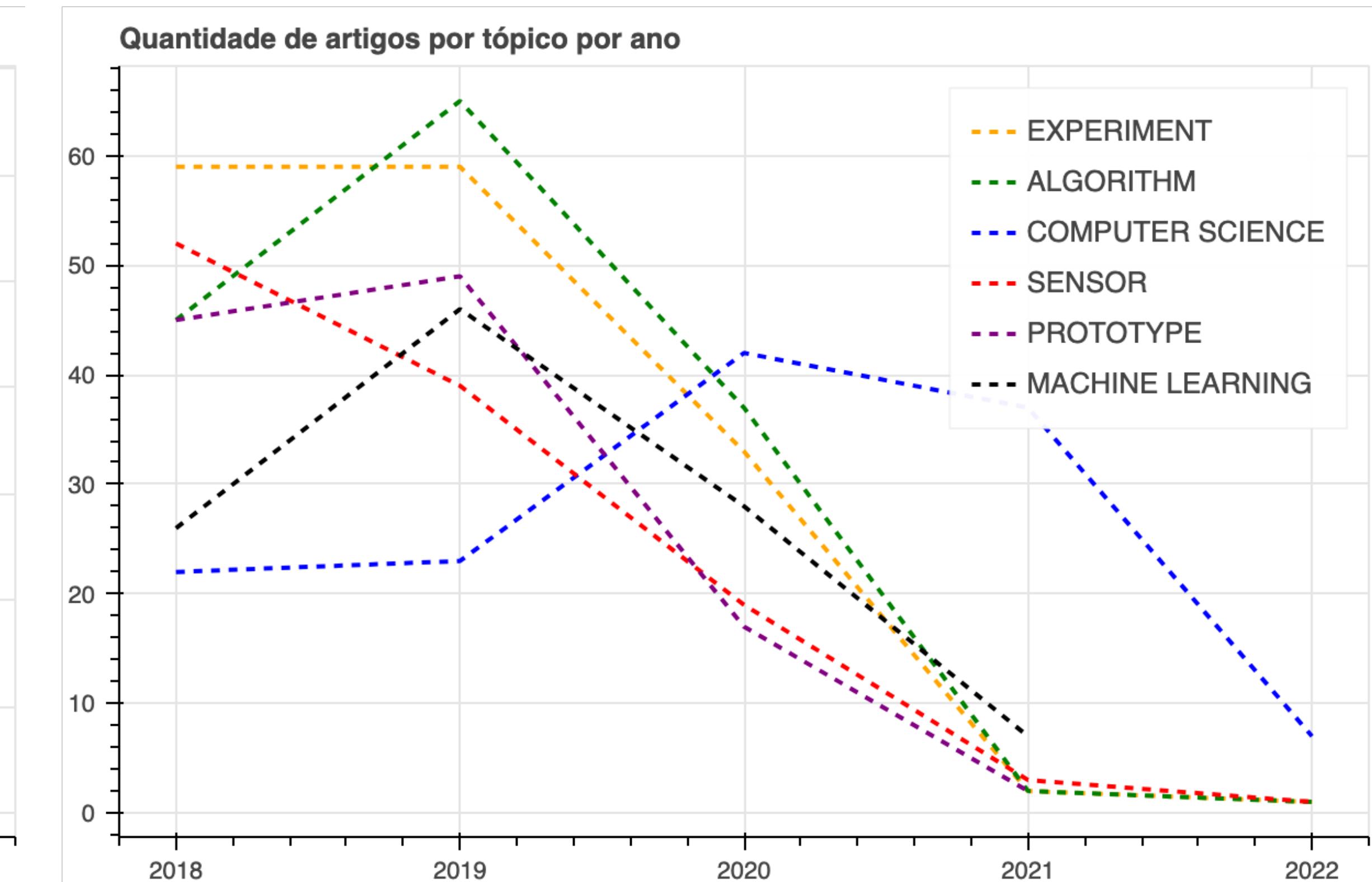
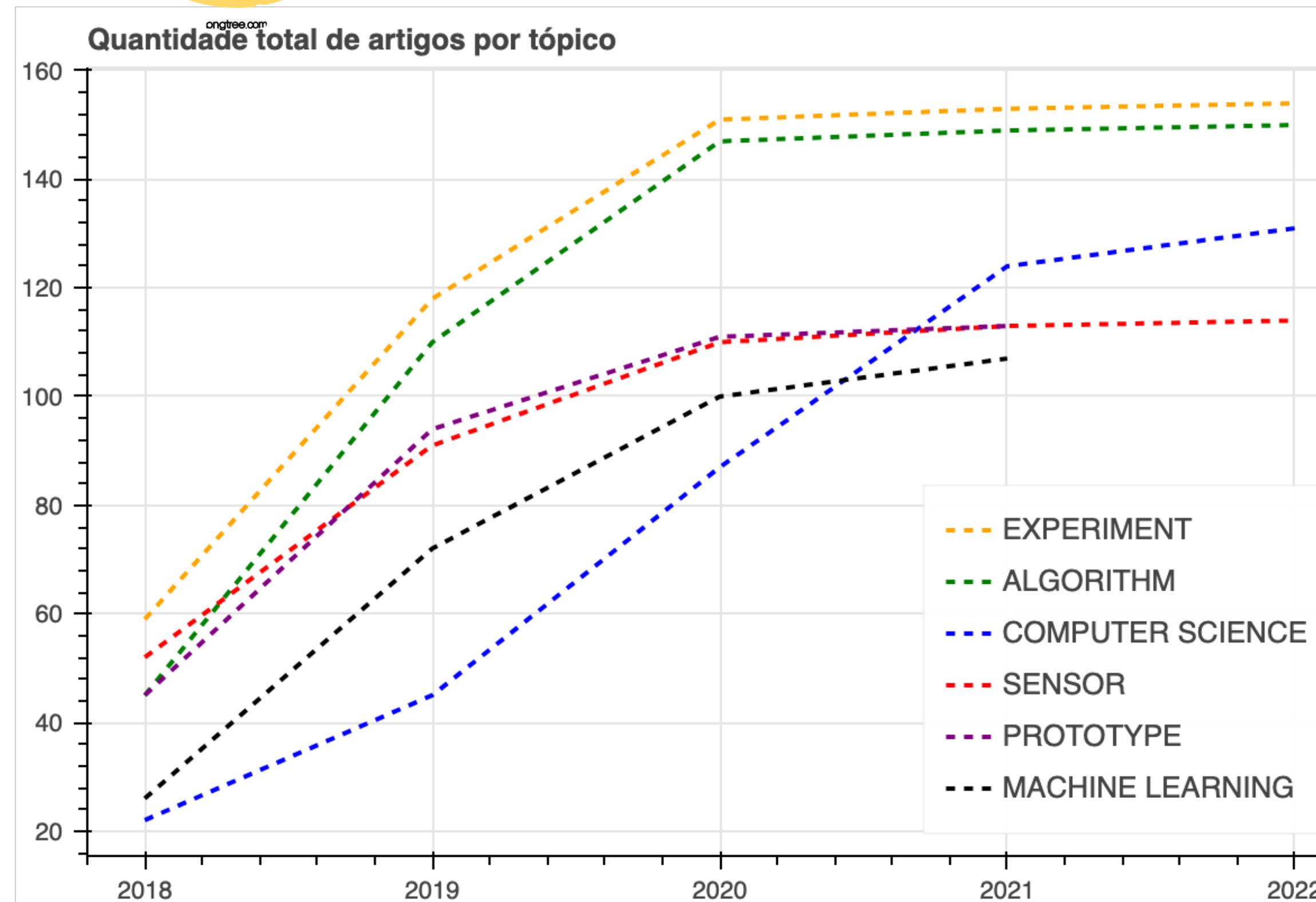


MACHINE LEARNING

Análises e resultados



Como foram as evoluções ao longo dos anos de publicações das conferências para cada tópico?



Análises e resultados



Para cada tópico, em que categoria as conferências entram em relação a publicação dos mesmos?

Dicionário de publicações de conferências por tópico

```
1 {  
2   'SOSP' : {  
3     'COMPUTER SYSTEMS ORGANIZATION': 23,  
4     'DEPENDABLE AND FAULT-TOLERANT SYSTEMS AND NETWORKS'  
5       : 18,  
6     'SECURITY AND PRIVACY': 12,  
7     'DATABASE AND STORAGE SECURITY': 2  
8     ...  
9 }
```

Análises e resultados



Para cada tópico, em que categoria as conferências entram em relação a publicação dos mesmos?

Dicionário de publicações de conferências por ano

```
1 {  
2   'SOSP': {  
3     2021: 54,  
4     2019: 38  
5   },  
6   'OSDI': {  
7     2021: 31,  
8     2020: 70,  
9     2018: 47  
10  }  
11 }
```

Análises e resultados



Para cada tópico, em que categoria as conferências entram em relação a publicação dos mesmos?

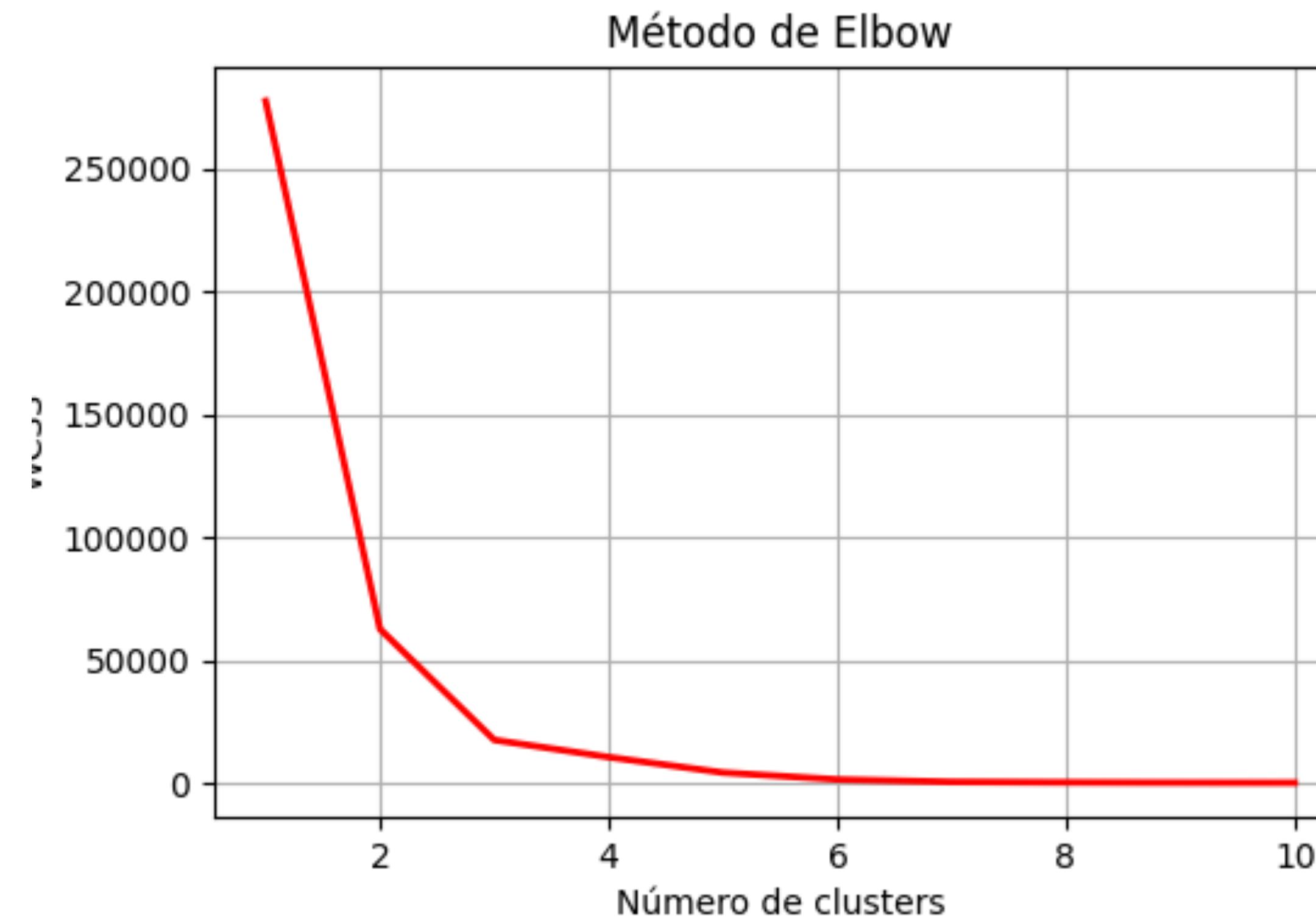
Para cada um dos 6 tópicos: Pega a quantidade total de publicações por conferência e compara com a quantidade de publicações do tópico na conferência

Algoritmo de clusterização K-Means por tópico
(Quantidade total do evento vs quantidade do evento pro tópico)

Análises e resultados



Para cada tópico, em que categoria as conferências entram em relação a publicação dos mesmos?

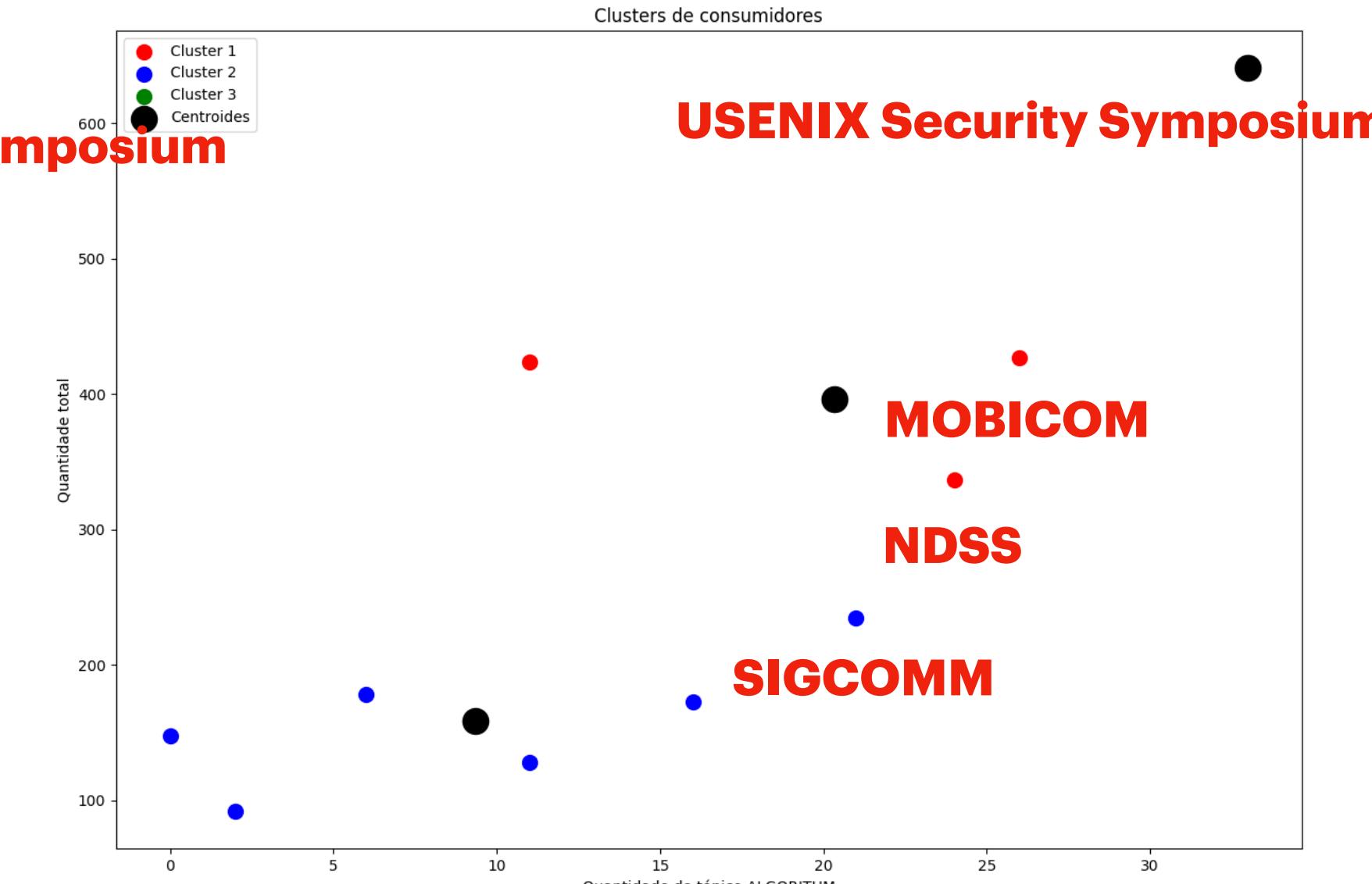
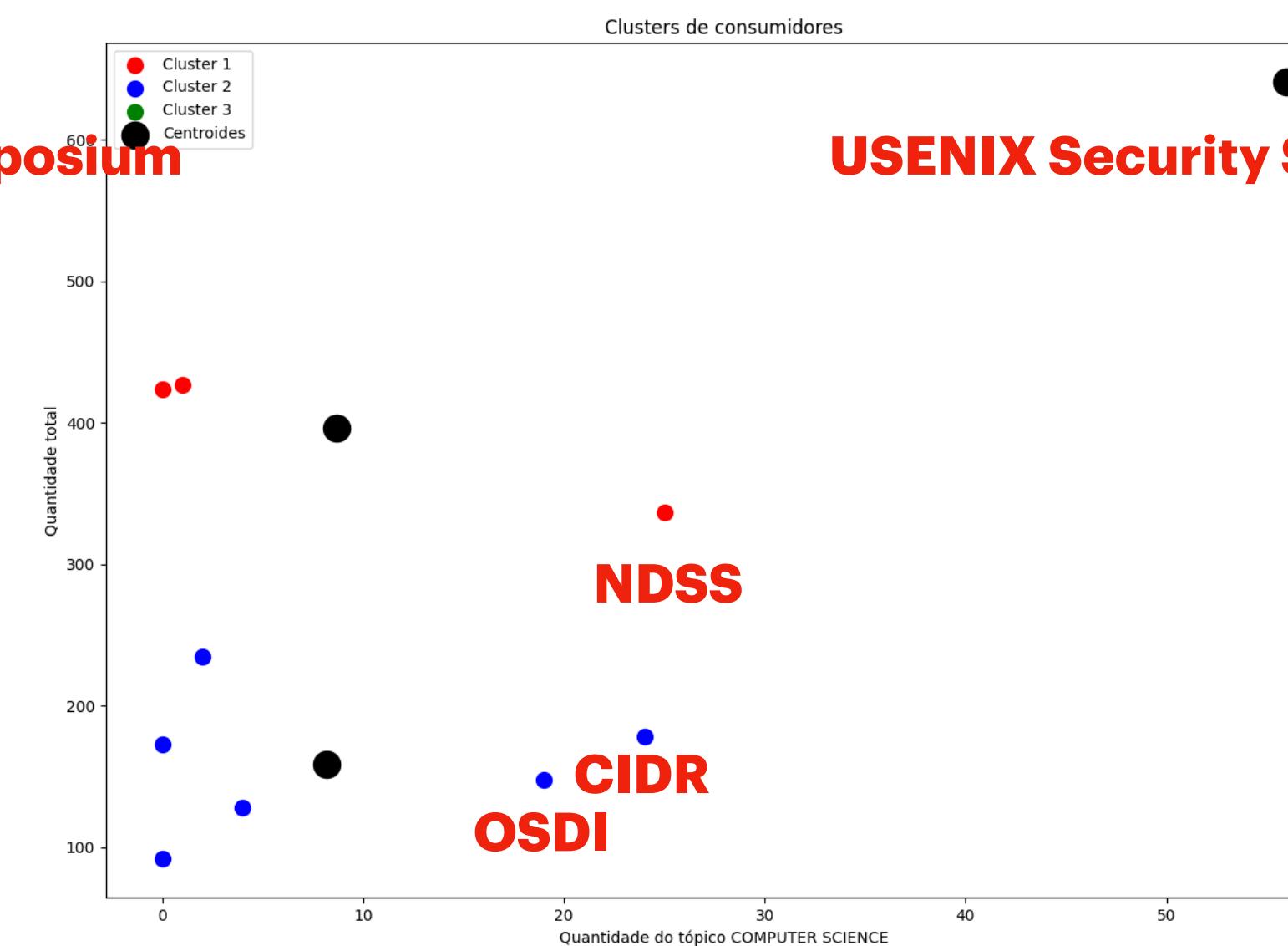
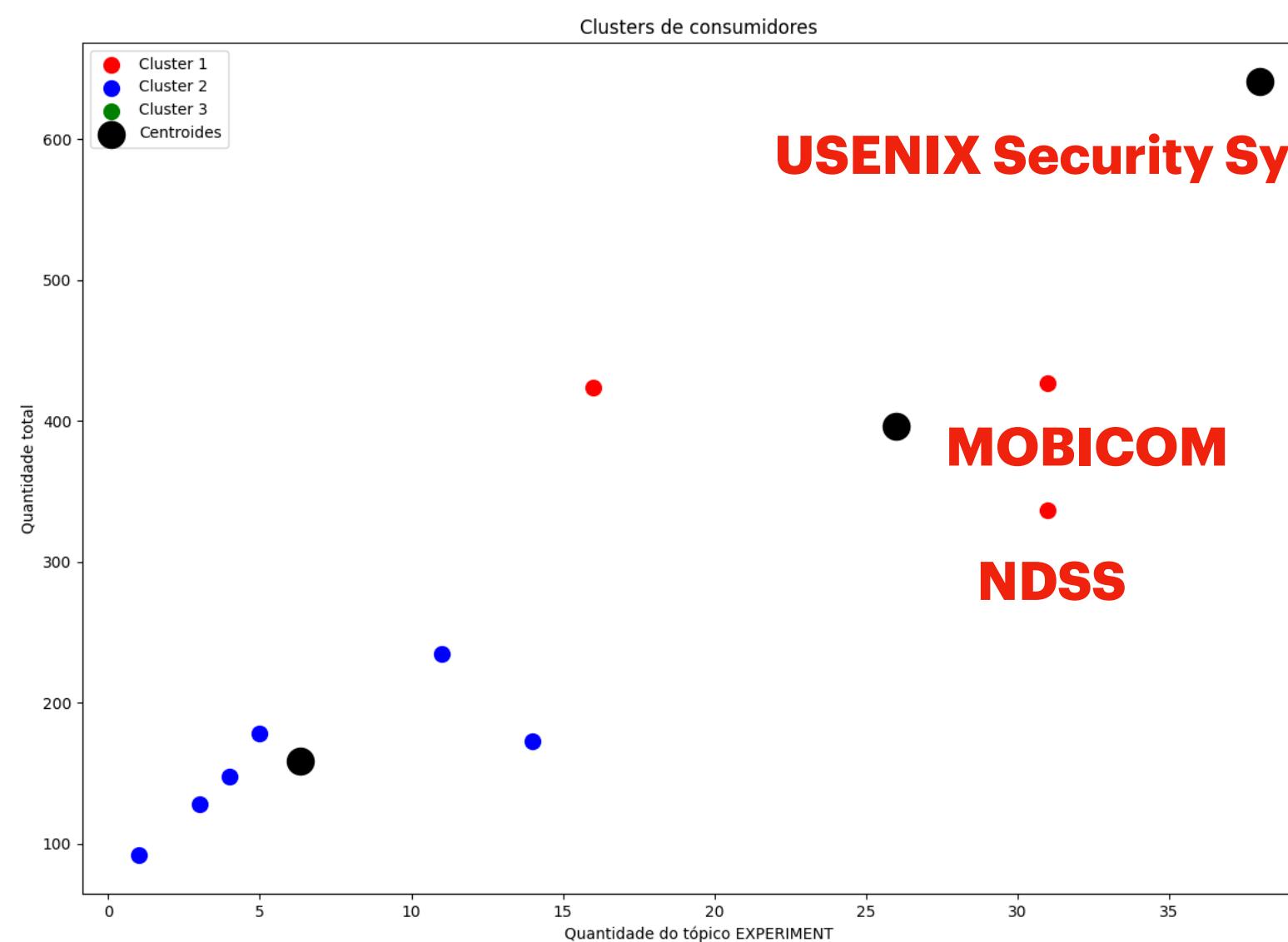


Quantidade
de Clusters: 3

Análises e resultados



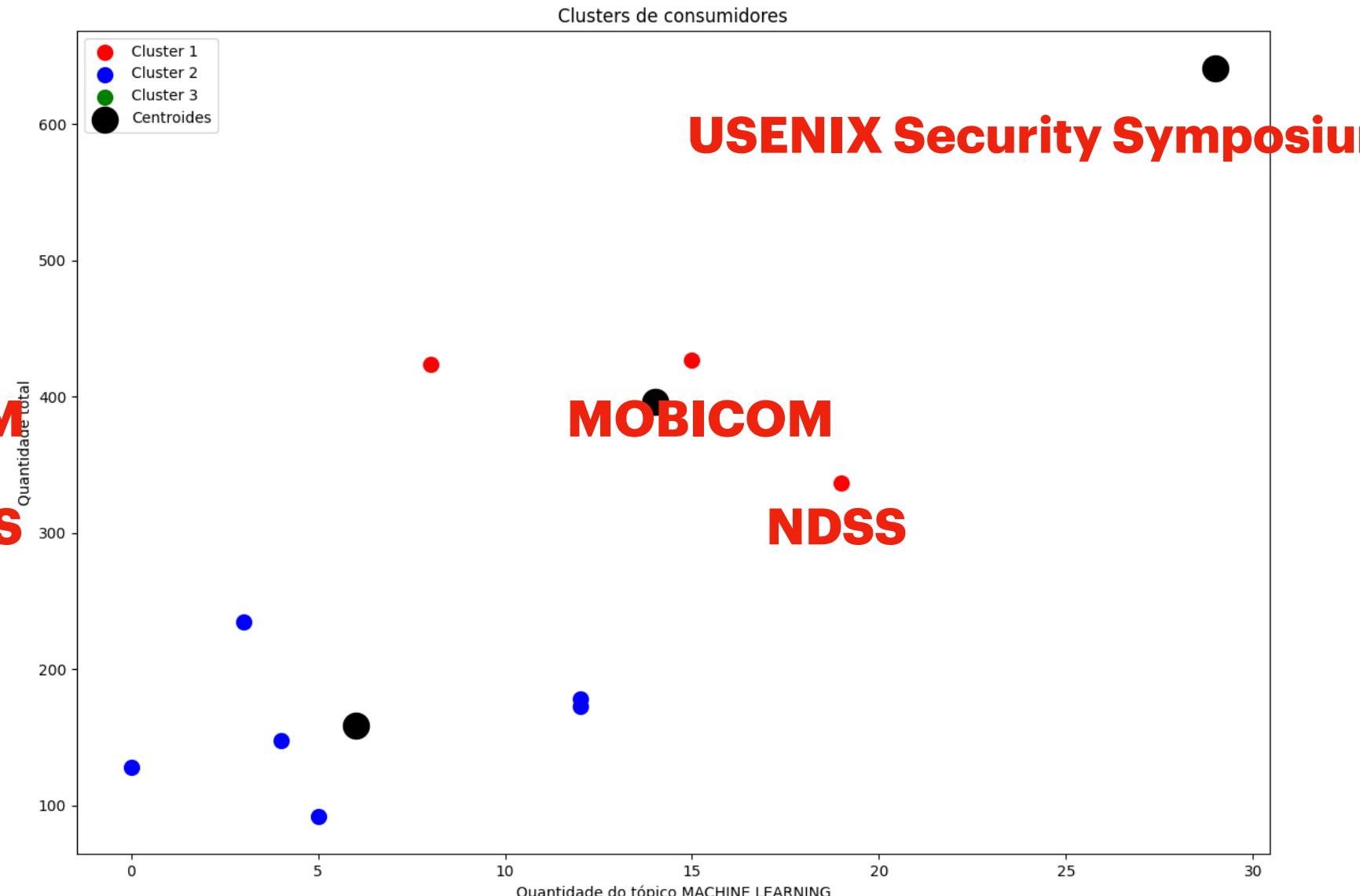
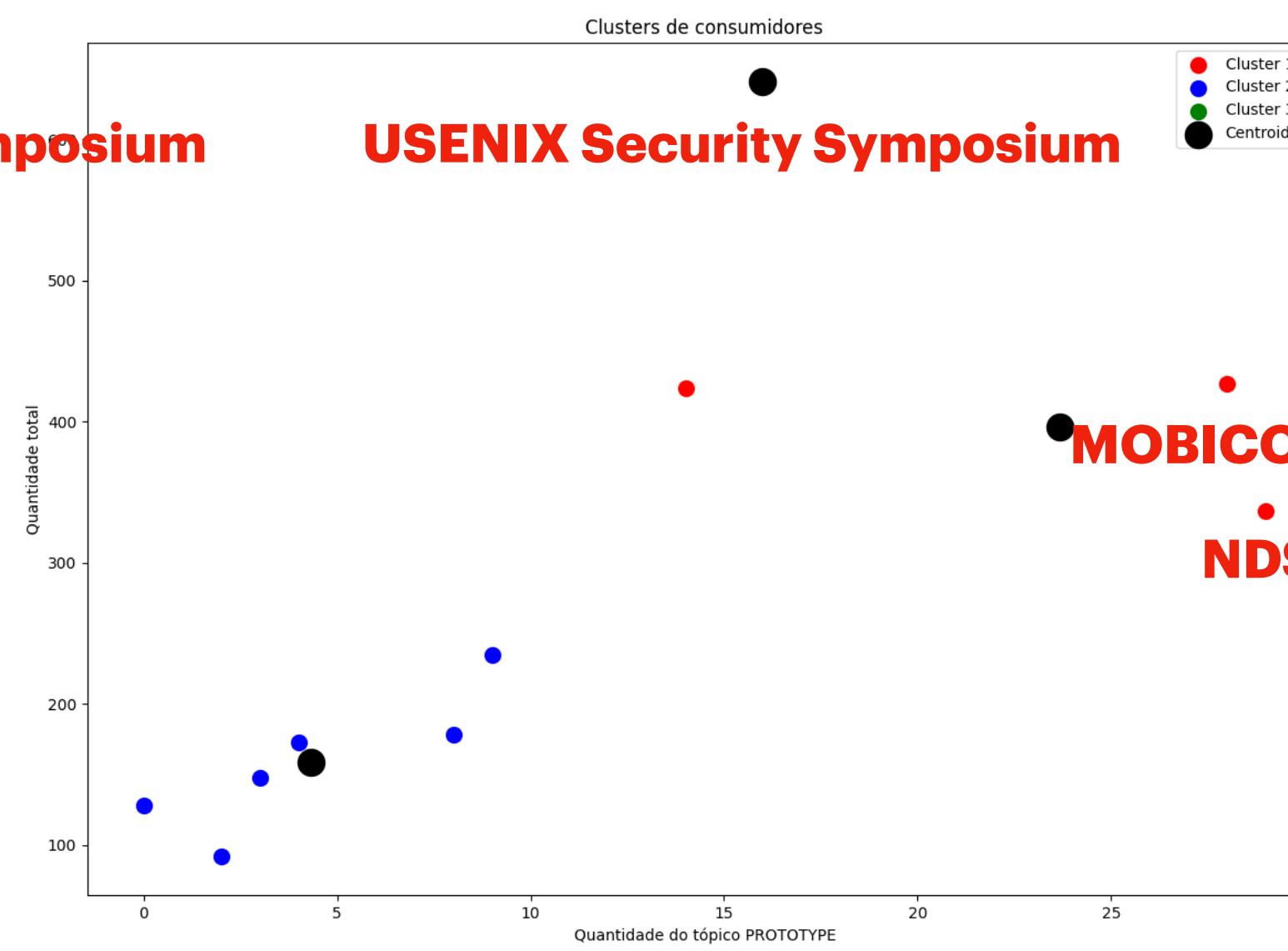
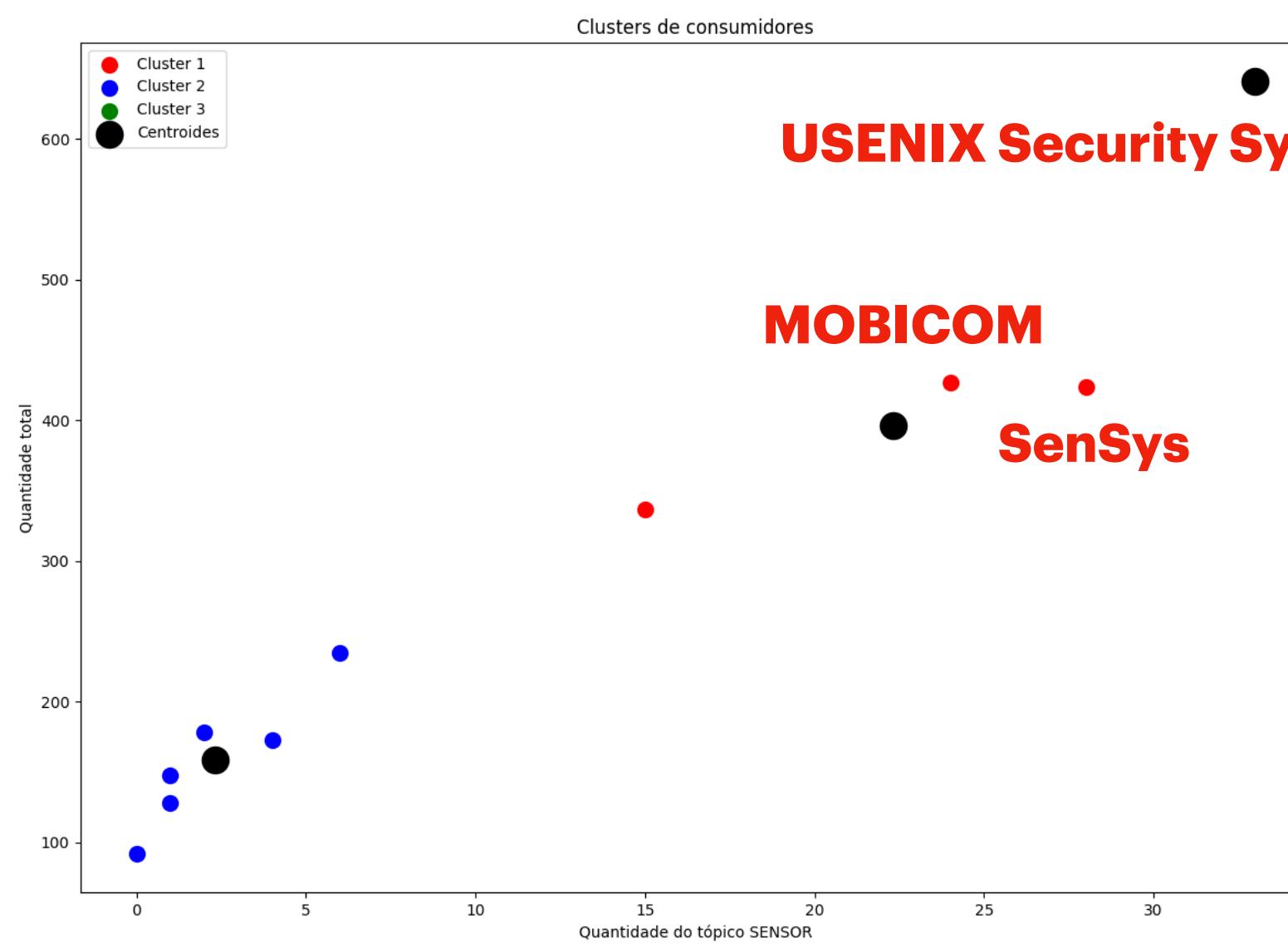
Para cada tópico, em que categoria as conferências entram em relação a publicação dos mesmos?



Análises e resultados



Para cada tópico, em que categoria as conferências entram em relação a publicação dos mesmos?



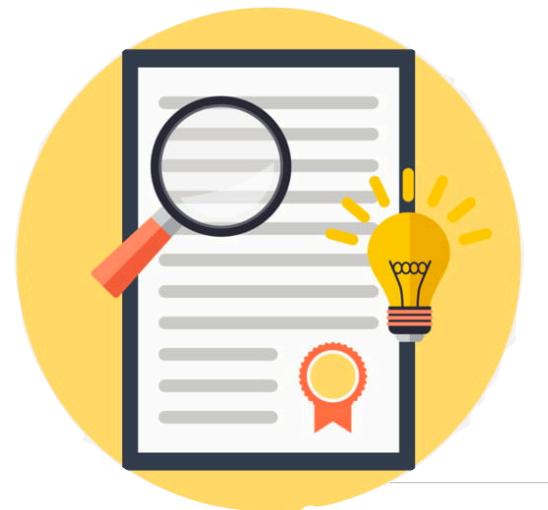
Análises e resultados



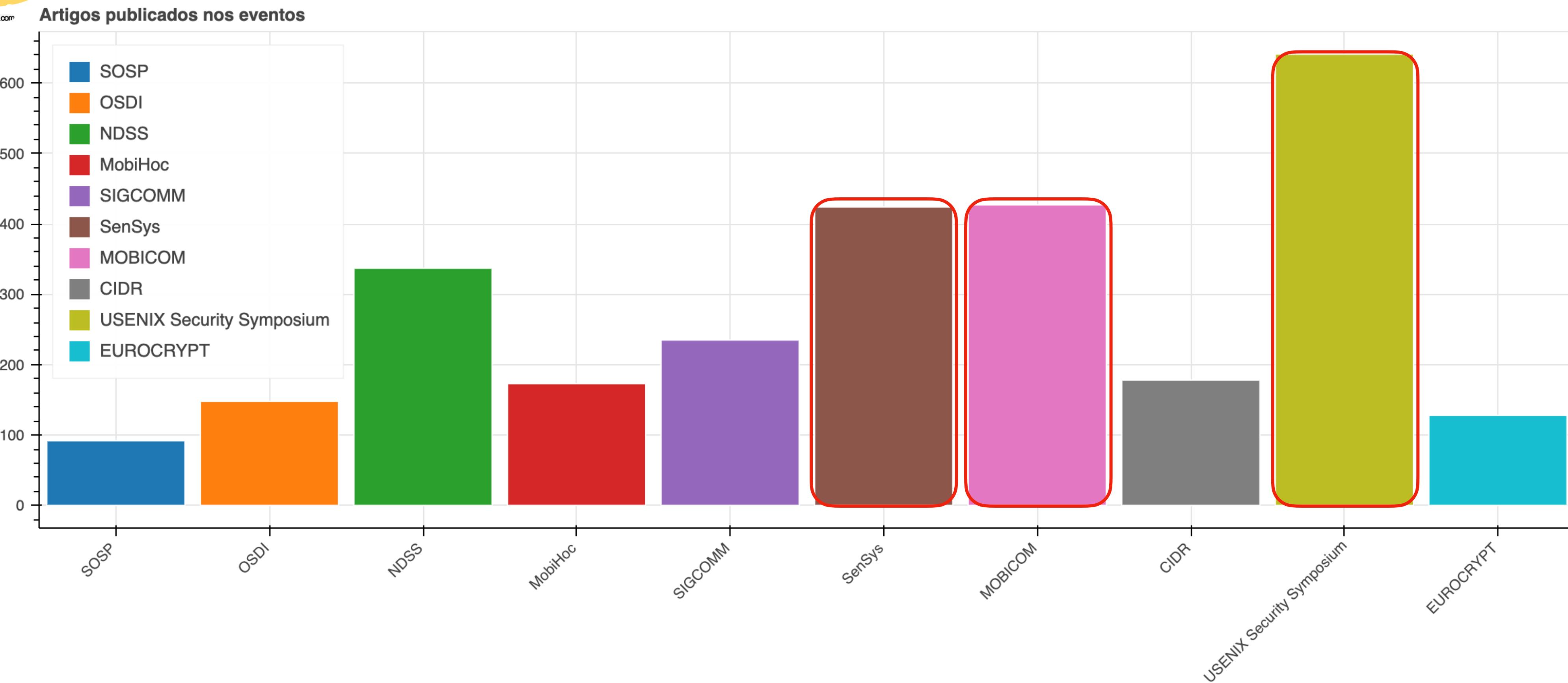
Quais as conferências que mais publicaram?

Utiliza o dicionário de publicações de conferências por ano e faz a soma de publicações por conferência - Total de publicações

Análises e resultados



Quais as conferências que mais publicaram?



Análises e resultados



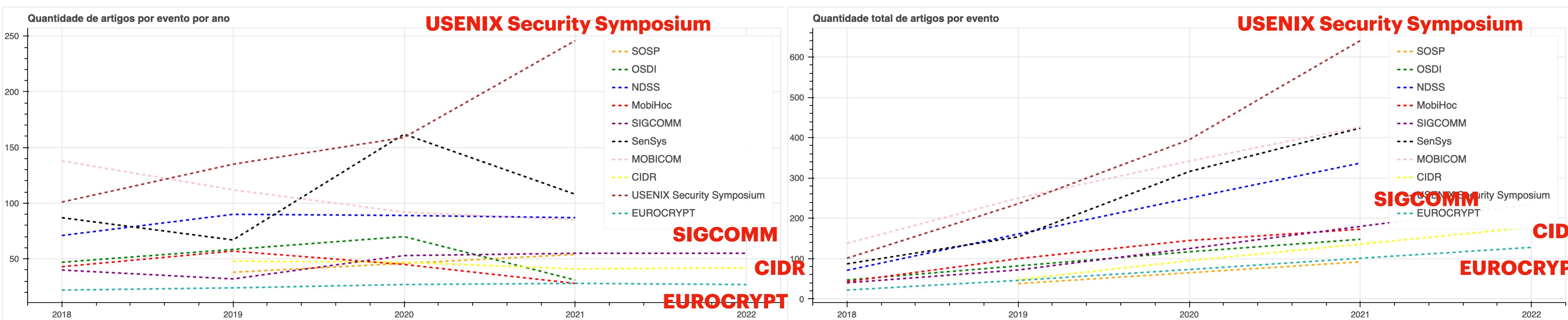
Como foram as evoluções ao longo dos anos para as conferências?

Utiliza o dicionário de publicações de conferências por ano e vai somando a quantidade de publicações por conferência ao longo dos anos

Análises e resultados



Como foram as evoluções ao longo dos anos para as conferências?



Análises e resultados

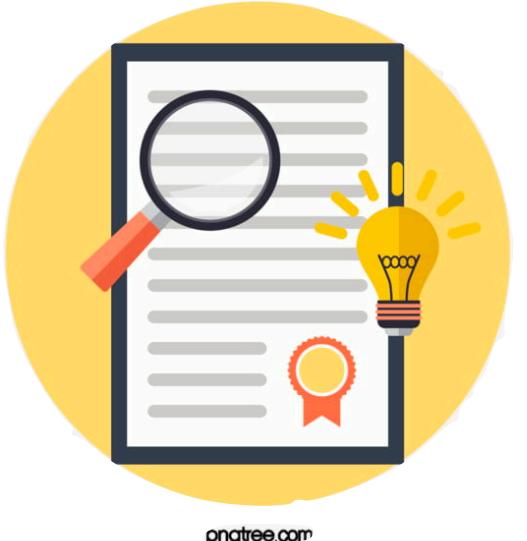


Para cada conferência, qual o tópico que mais aparece?

Para cada conferência, pega os 6 tópicos que mais tem publicações

Compara a quantidade total de publicações do tópico com a quantidade de publicações dele na conferência

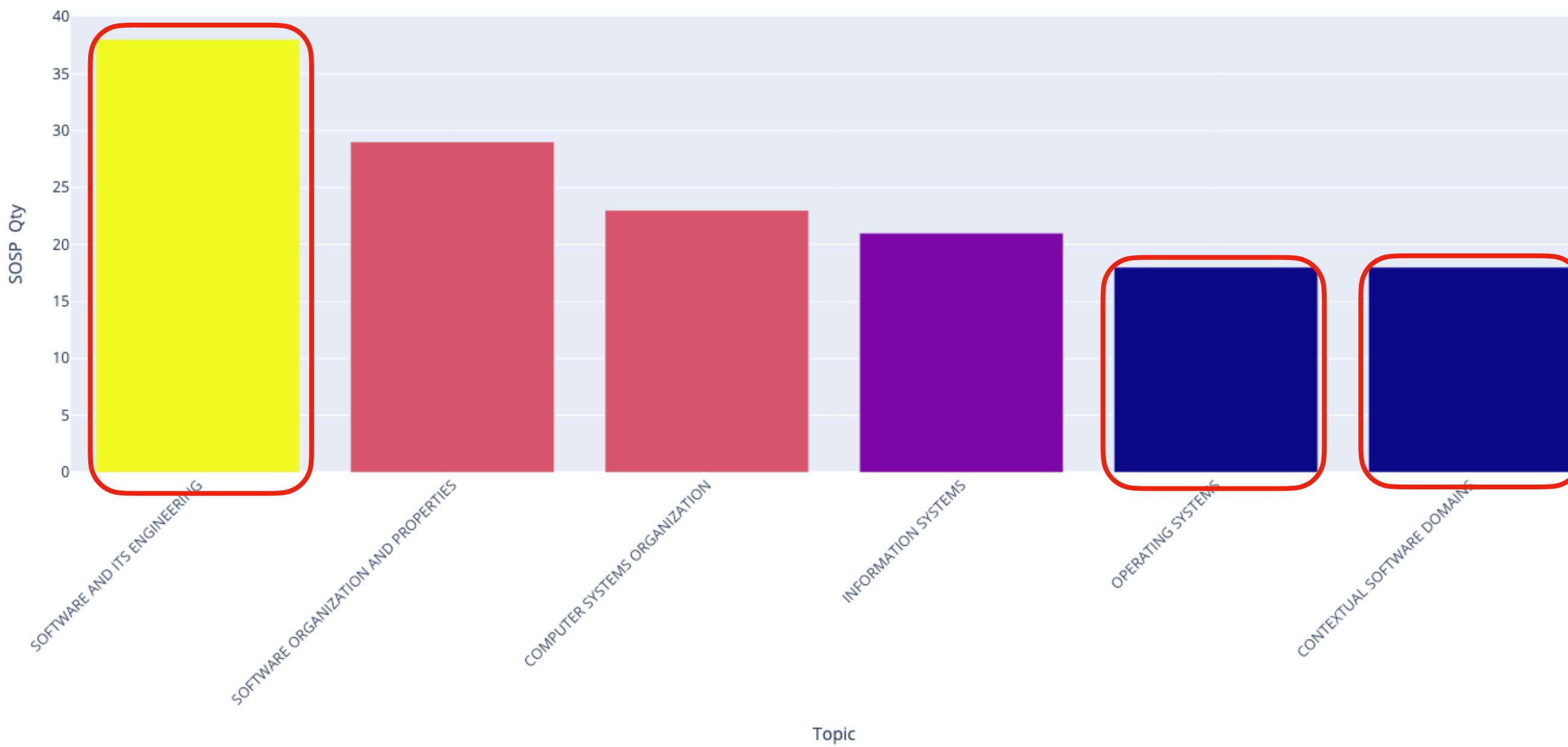
Análises e resultados



Para cada conferência, qual o tópico que mais aparece?

SOSP

Quantidade de publicações dos tópicos no SOSP



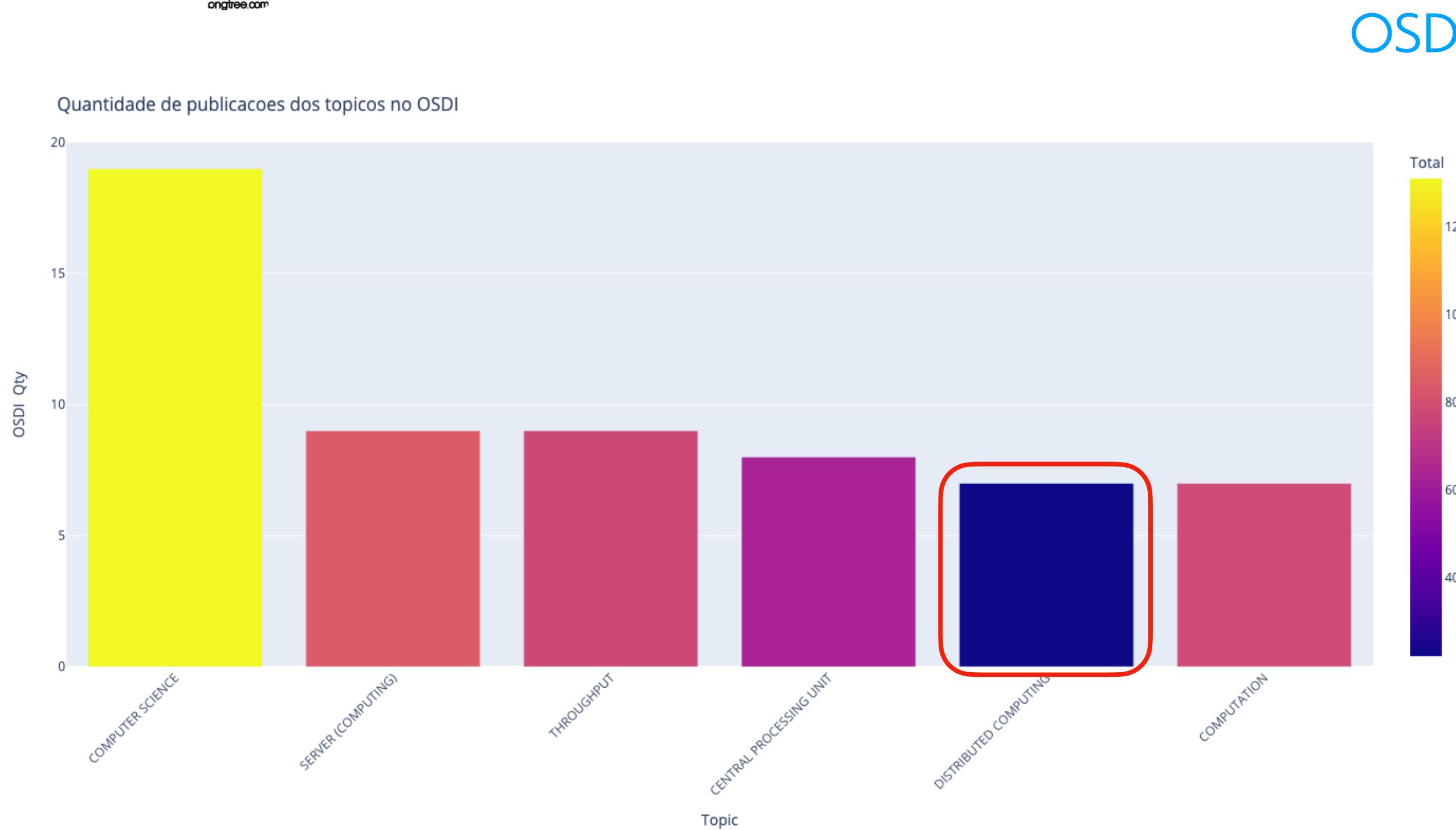
A table titled "SOSP" showing the total number of publications ("Total") and the count for the SOSP conference ("SOSP Qty"). The table includes a color scale legend on the left.

Topic	Total	SOSP	Qty
SOFTWARE AND ITS ENGINEERING	38	38.0	
SOFTWARE ORGANIZATION AND PROPERTIES	29	29.0	
COMPUTER SYSTEMS ORGANIZATION	23	23.0	
INFORMATION SYSTEMS	18	18.0	
OPERATING SYSTEMS	18	18.0	
CONTEXTUAL SOFTWARE DOMAINS	18	18.0	

Análises e resultados



Para cada conferência, qual o tópico que mais aparece?



Topic	Total	OSDI	Qty
COMPUTER SCIENCE	131	19.0	19.0
SERVER (COMPUTING)	84	9.0	9.0
THROUGHPUT	77	9.0	9.0
CENTRAL PROCESSING UNIT	63	8.0	8.0
DISTRIBUTED COMPUTING	22	7.0	7.0
COMPUTATION	78	7.0	7.0

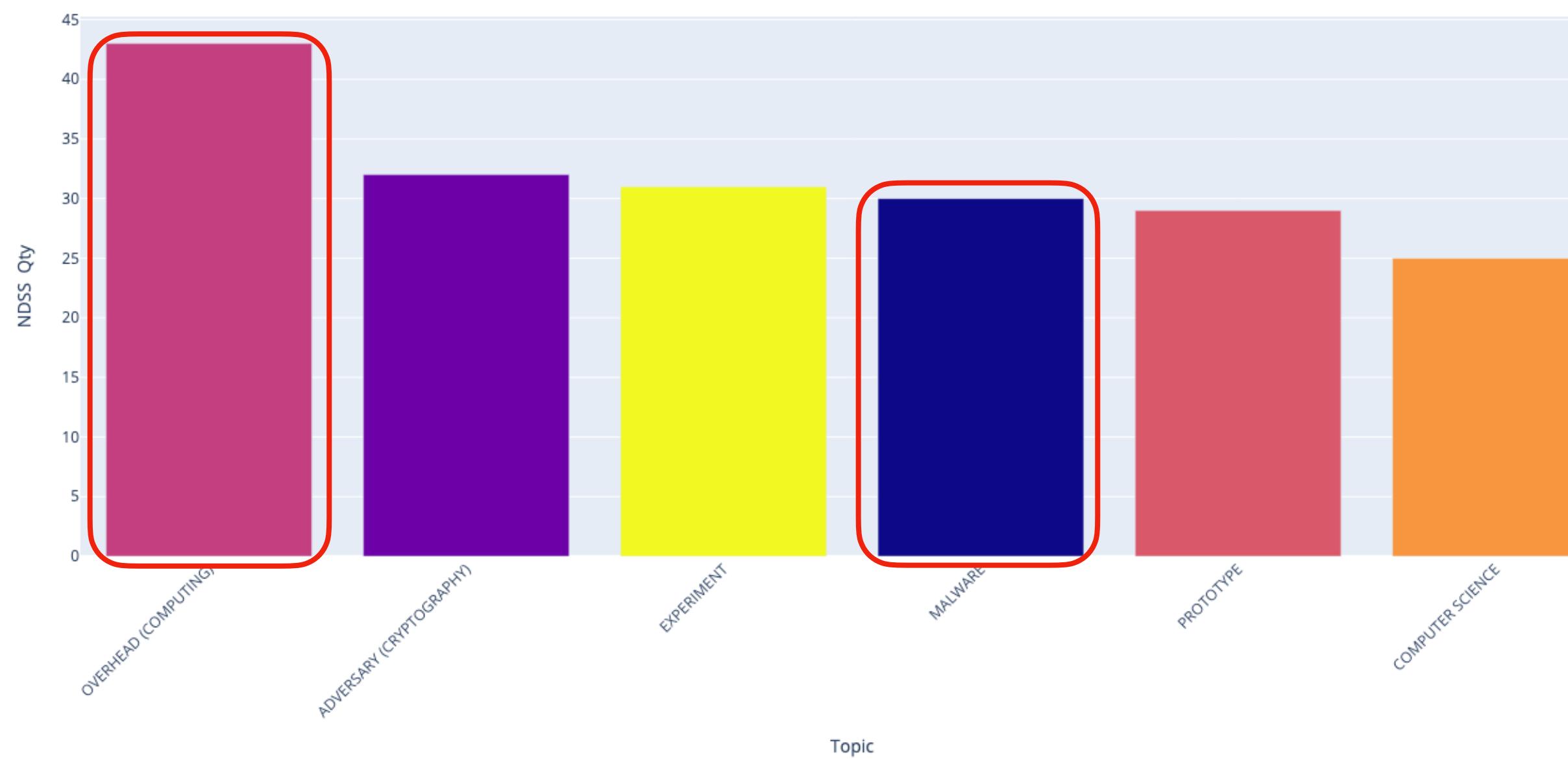
Análises e resultados



Para cada conferência, qual o tópico que mais aparece?

NDSS

Quantidade de publicações dos tópicos no NDSS



Topic	Total	NDSS	Qty
0 OVERHEAD (COMPUTING)	105		43.0
1 ADVERSARY (CRYPTOGRAPHY)	80		32.0
2 EXPERIMENT	154		31.0
3 MALWARE	61		30.0
4 PROTOTYPE	113		29.0
5 COMPUTER SCIENCE	131		25.0

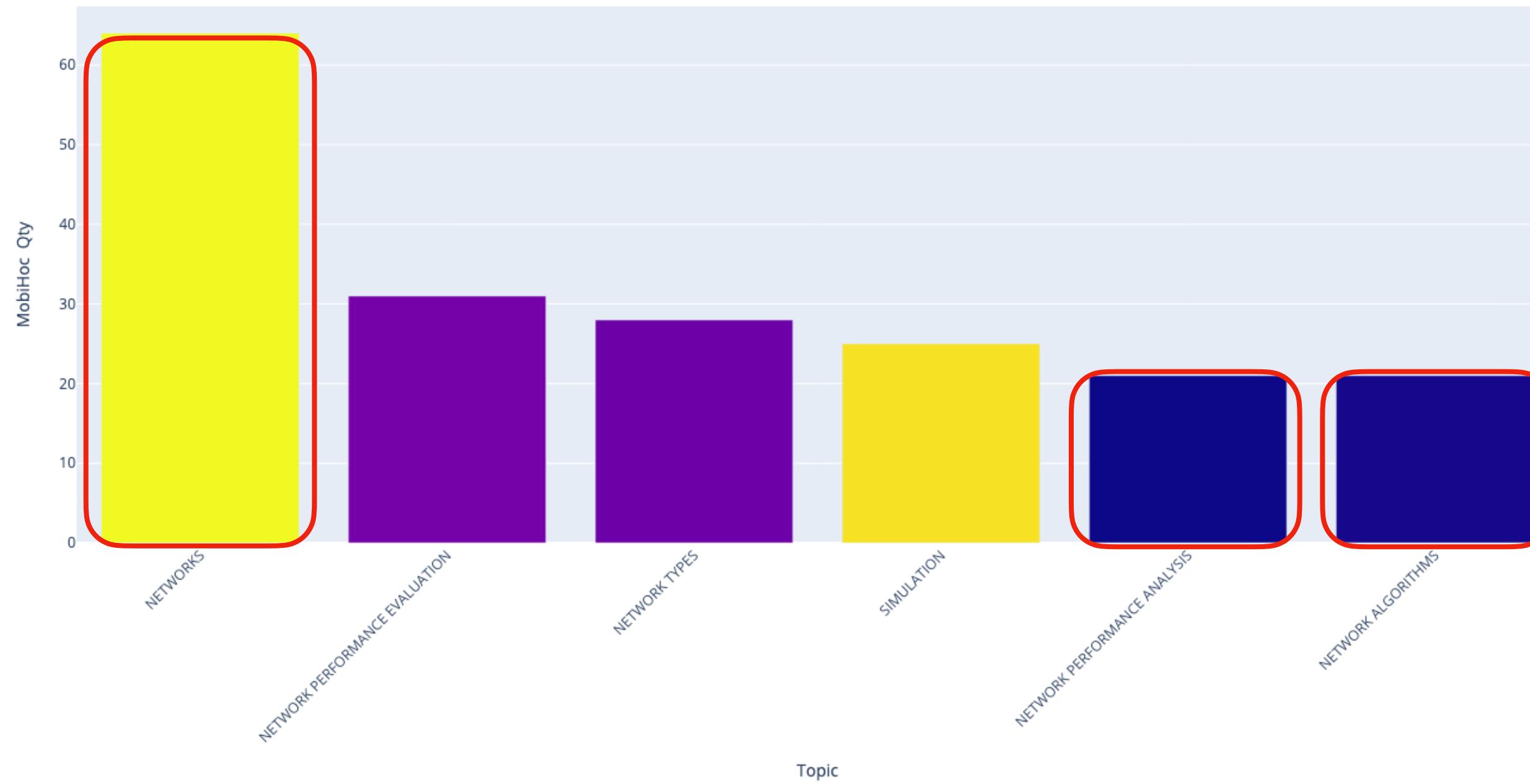
Análises e resultados



Para cada conferência, qual o tópico que mais aparece?

MobiHoc

Quantidade de publicações dos tópicos no MobiHoc



A vertical color scale legend titled "Total" with a gradient from dark purple at the bottom to bright yellow at the top, with numerical markers at 30, 40, 50, 60, and 70.

Topic	Total	MobiHoc	Qty
NETWORKS	74	64.0	
NETWORK PERFORMANCE EVALUATION	33	31.0	
NETWORK TYPES	32	28.0	
SIMULATION	71	25.0	
NETWORK PERFORMANCE ANALYSIS	21	21.0	
NETWORK ALGORITHMS	22	21.0	

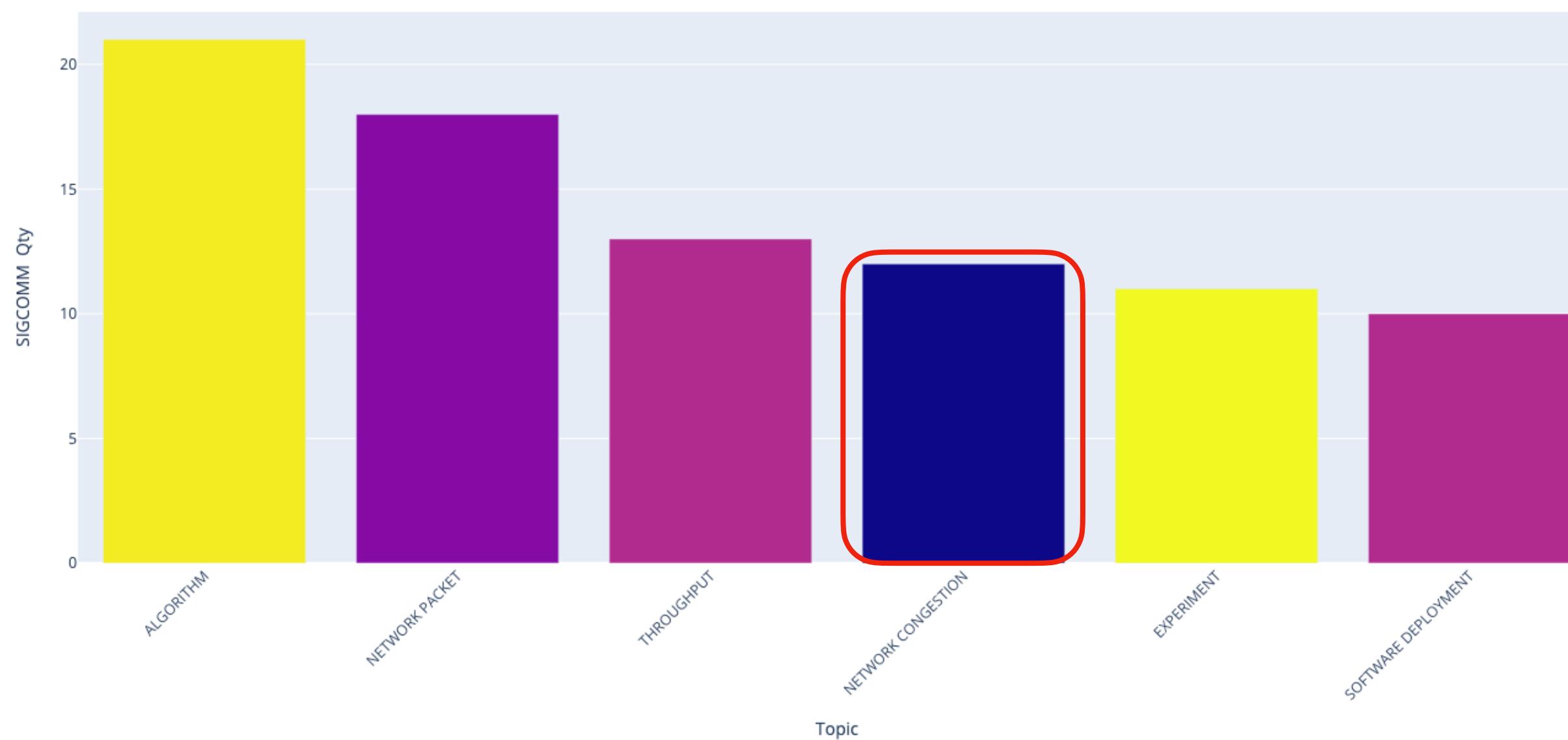
Análises e resultados



Para cada conferência, qual o tópico que mais aparece?

SIGCOMM

Quantidade de publicações dos tópicos no SIGCOMM



	Topic	Total	SIGCOMM	Qty
0	ALGORITHM	150		21.0
1	NETWORK PACKET	60		18.0
2	THROUGHPUT	77		13.0
3	NETWORK CONGESTION	25		12.0
4	EXPERIMENT	154		11.0
5	SOFTWARE DEPLOYMENT	77		10.0

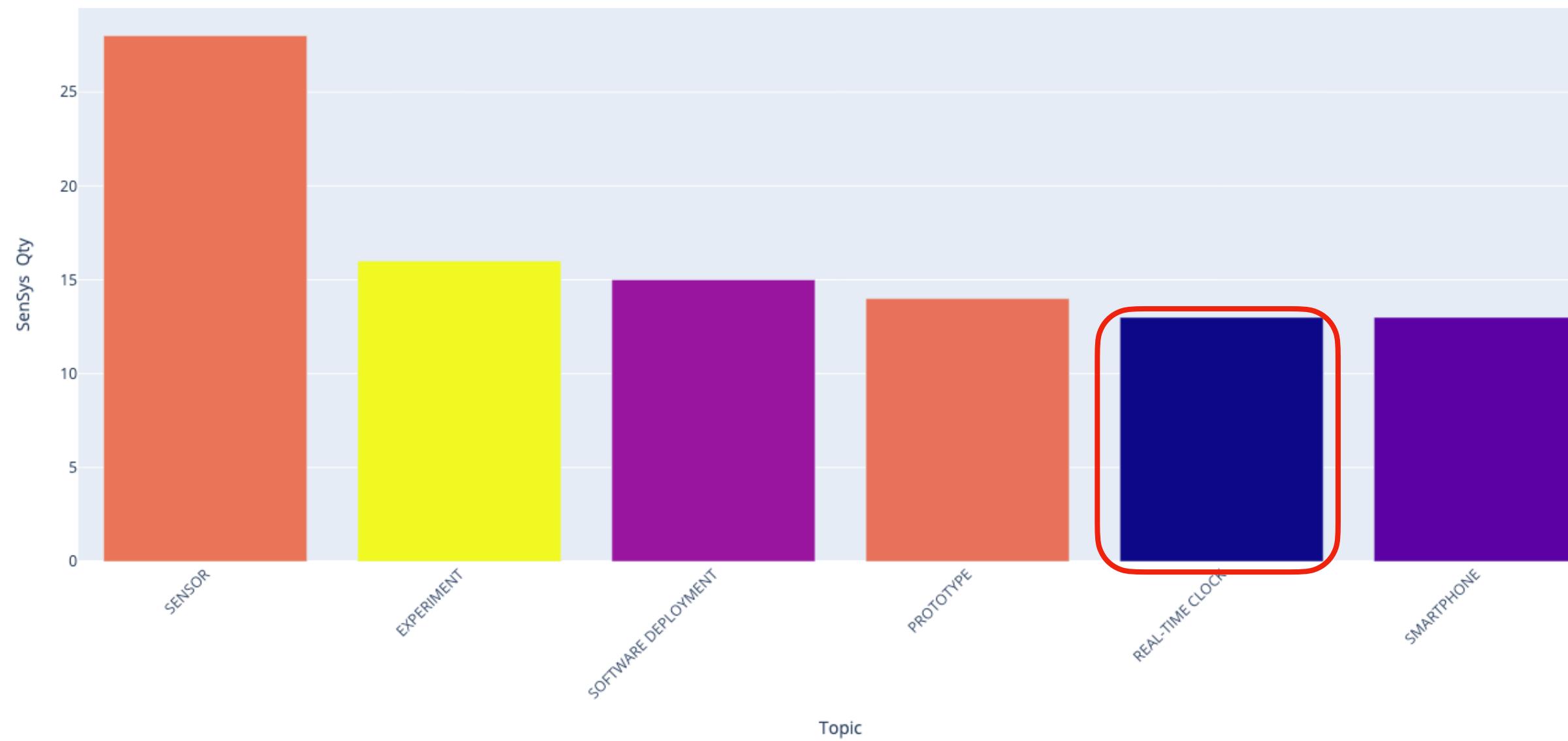
Análises e resultados



Para cada conferência, qual o tópico que mais aparece?

SenSys

Quantidade de publicações dos tópicos no SenSys

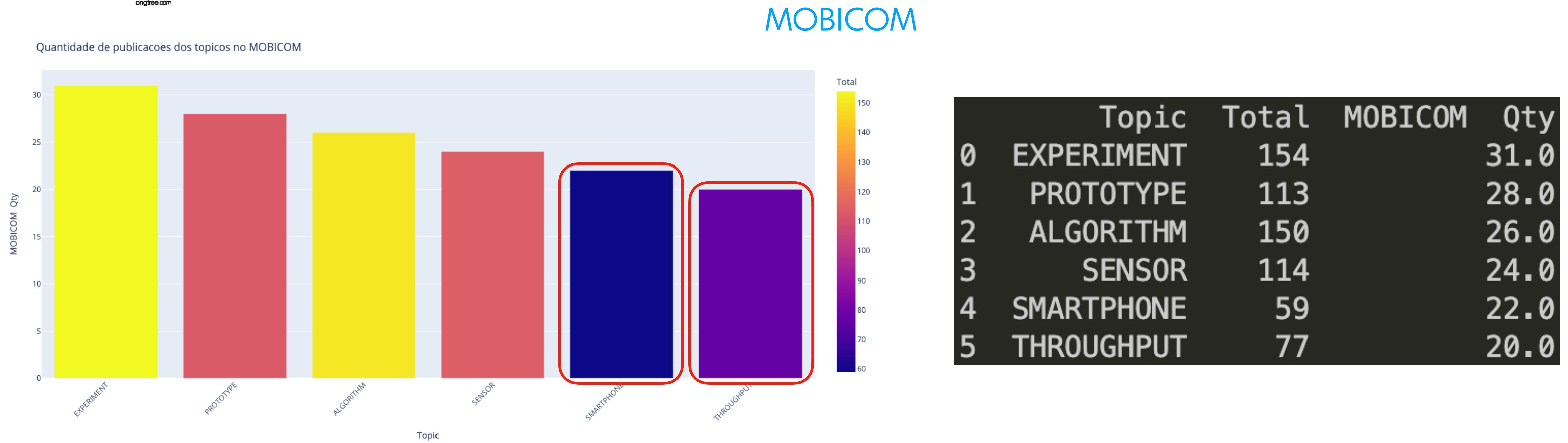


Total	Topic	Total	SenSys	Qty
0	SENSOR	114		28.0
1	EXPERIMENT	154		16.0
2	SOFTWARE DEPLOYMENT	77		15.0
3	PROTOTYPE	113		14.0
4	REAL-TIME CLOCK	40		13.0
5	SMARTPHONE	59		13.0

Análises e resultados



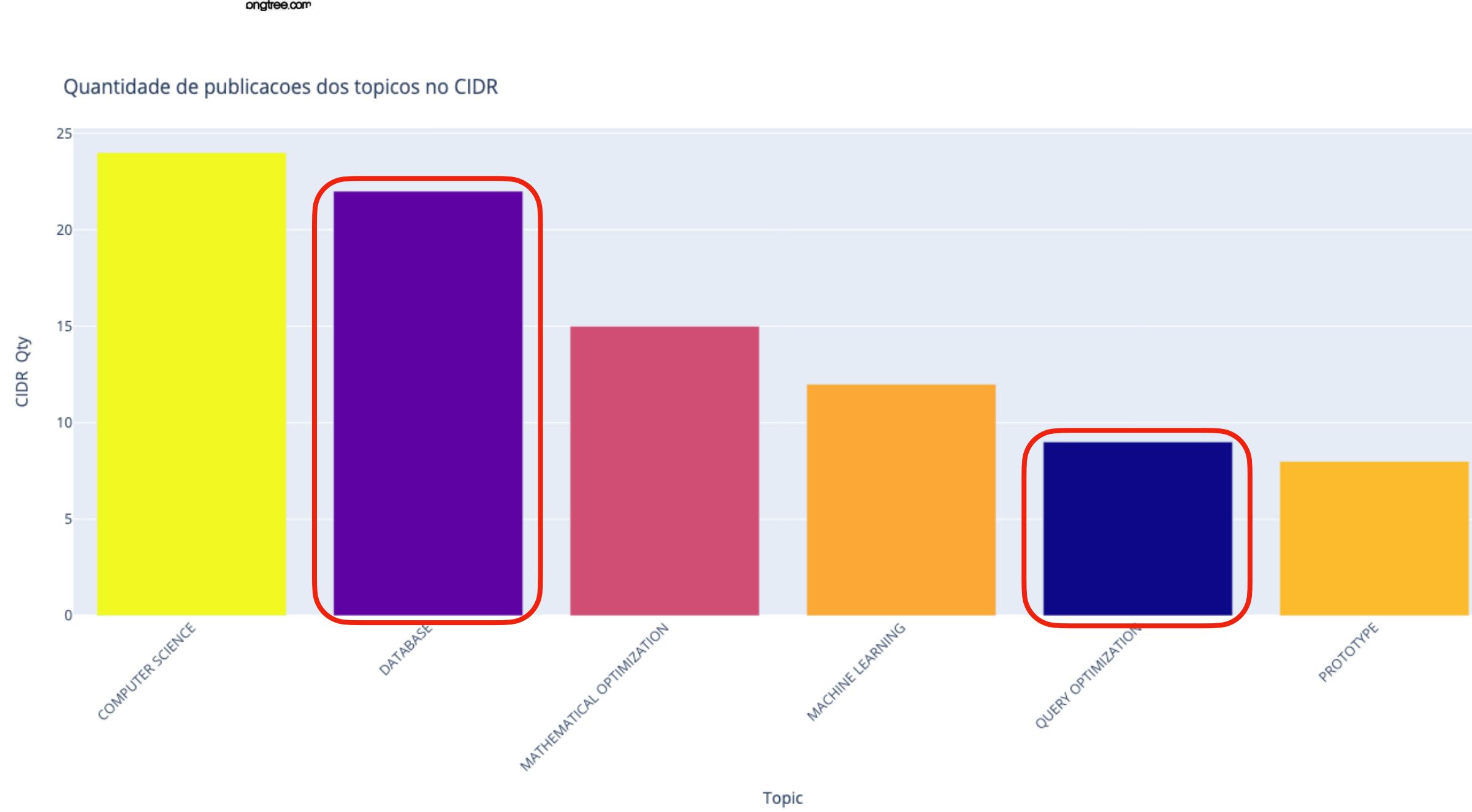
Para cada conferência, qual o tópico que mais aparece?



Análises e resultados



Para cada conferência, qual o tópico que mais aparece?



Total

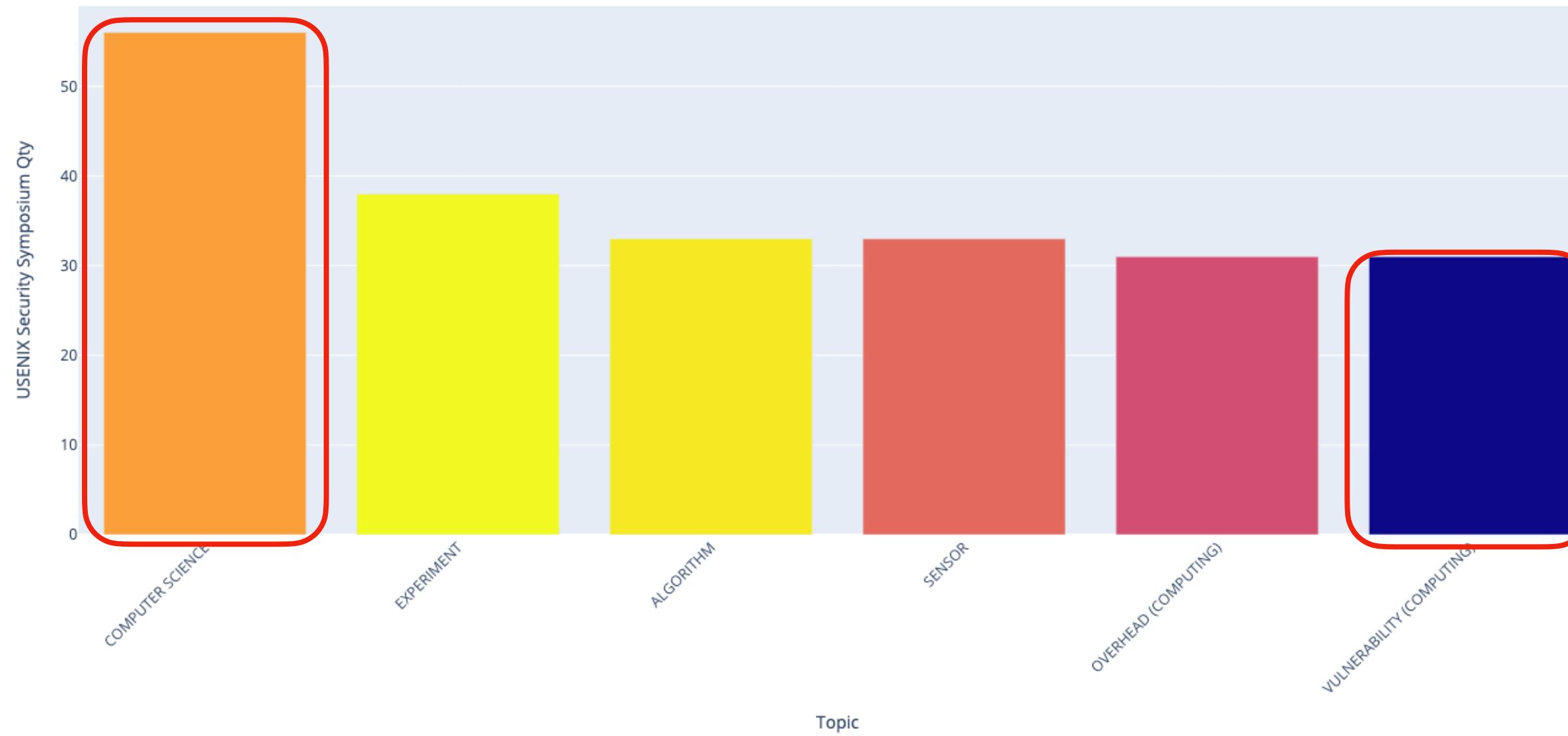
Topic	Total	CIDR	Qty
COMPUTER SCIENCE	131	24.0	24.0
DATABASE	31	22.0	22.0
MATHEMATICAL OPTIMIZATION	73	15.0	15.0
MACHINE LEARNING	107	12.0	12.0
QUERY OPTIMIZATION	10	9.0	9.0
PROTOTYPE	113	8.0	8.0

Análises e resultados



Para cada conferência, qual o tópico que mais aparece?

Quantidade de publicações dos tópicos no USENIX Security Symposium



USENIX Security Symposium

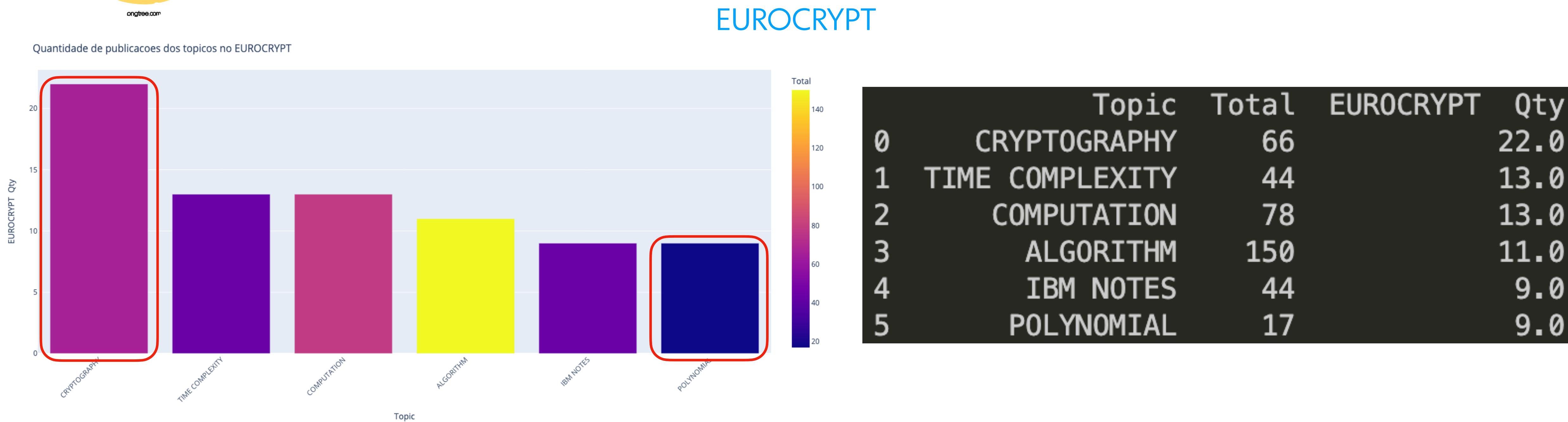


Topic	Total	USENIX Security Symposium Qty
COMPUTER SCIENCE	131	56.0
EXPERIMENT	154	38.0
ALGORITHM	150	33.0
SENSOR	114	33.0
OVERHEAD (COMPUTING)	105	31.0
VULNERABILITY (COMPUTING)	50	31.0

Análises e resultados



Para cada conferência, qual o tópico que mais aparece?



Análises e resultados



Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?

Dicionário de publicações de conferências por ano e tópico

```
1 {  
2   'MOBICOM' : {  
3     2020: {  
4       'LOCATION (GEOGRAPHY)' : 1,  
5       'DUPLEX (TELECOMMUNICATIONS)' : 1,  
6       'DEFINITION' : 1,  
7       'PHASED ARRAY' : 2,  
8       'EXPERIMENT' : 4,  
9       'FEEDBACK' : 1,  
10      'NETWORK PERFORMANCE' : 1,  
11      'APPROXIMATION ALGORITHM' : 1,  
12      'PROGRAMMING PARADIGM' : 1,  
13      'DOWNTIME' : 2,  
14      'IBM NOTES' : 1,  
15      'BACKSCATTER (EMAIL)' : 1,  
16      ...  
17    }  
18  }  
19 }
```

Análises e resultados



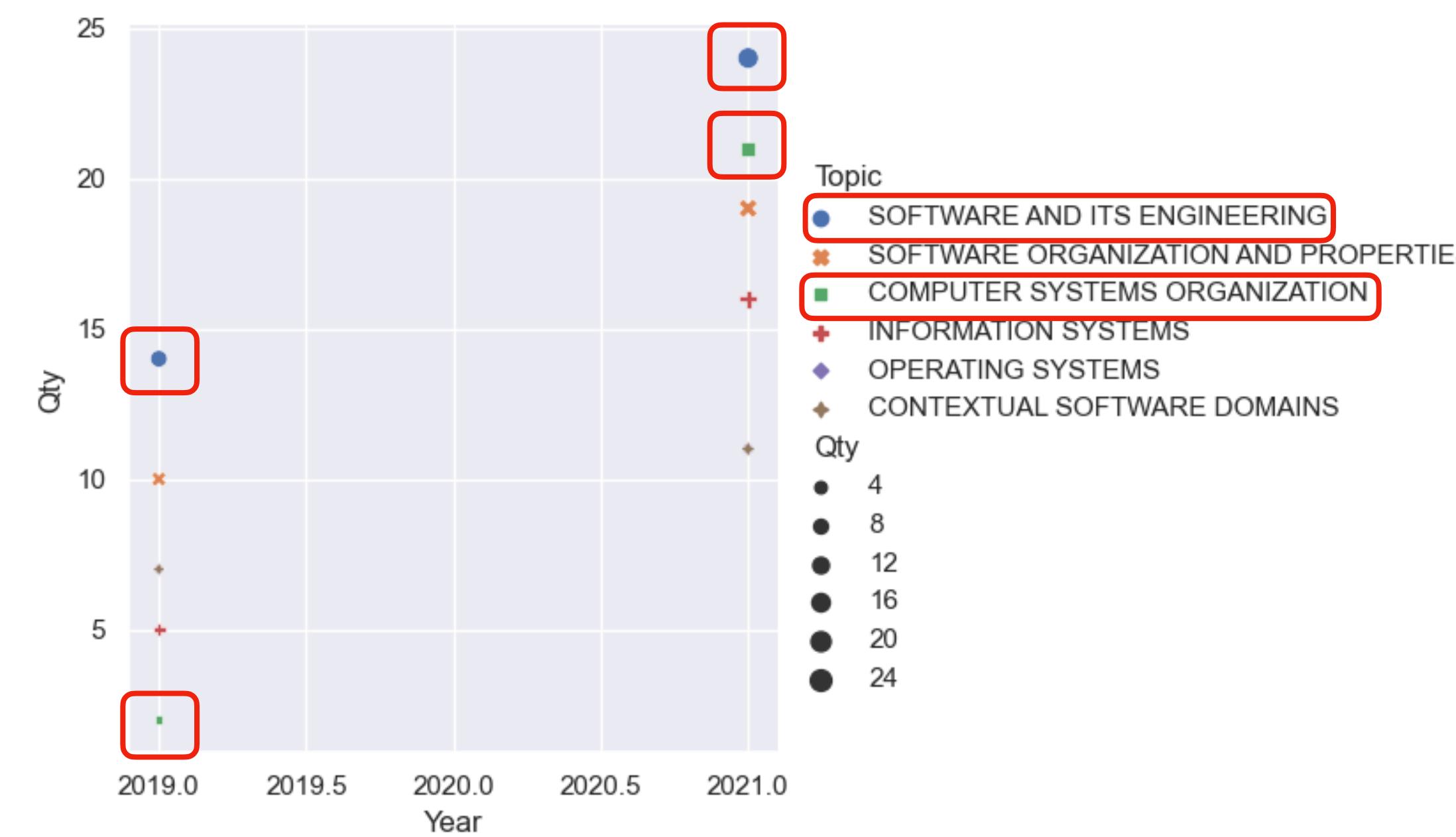
Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?

Pega os 6 tópicos com mais publicações para cada conferência e suas quantidades por ano

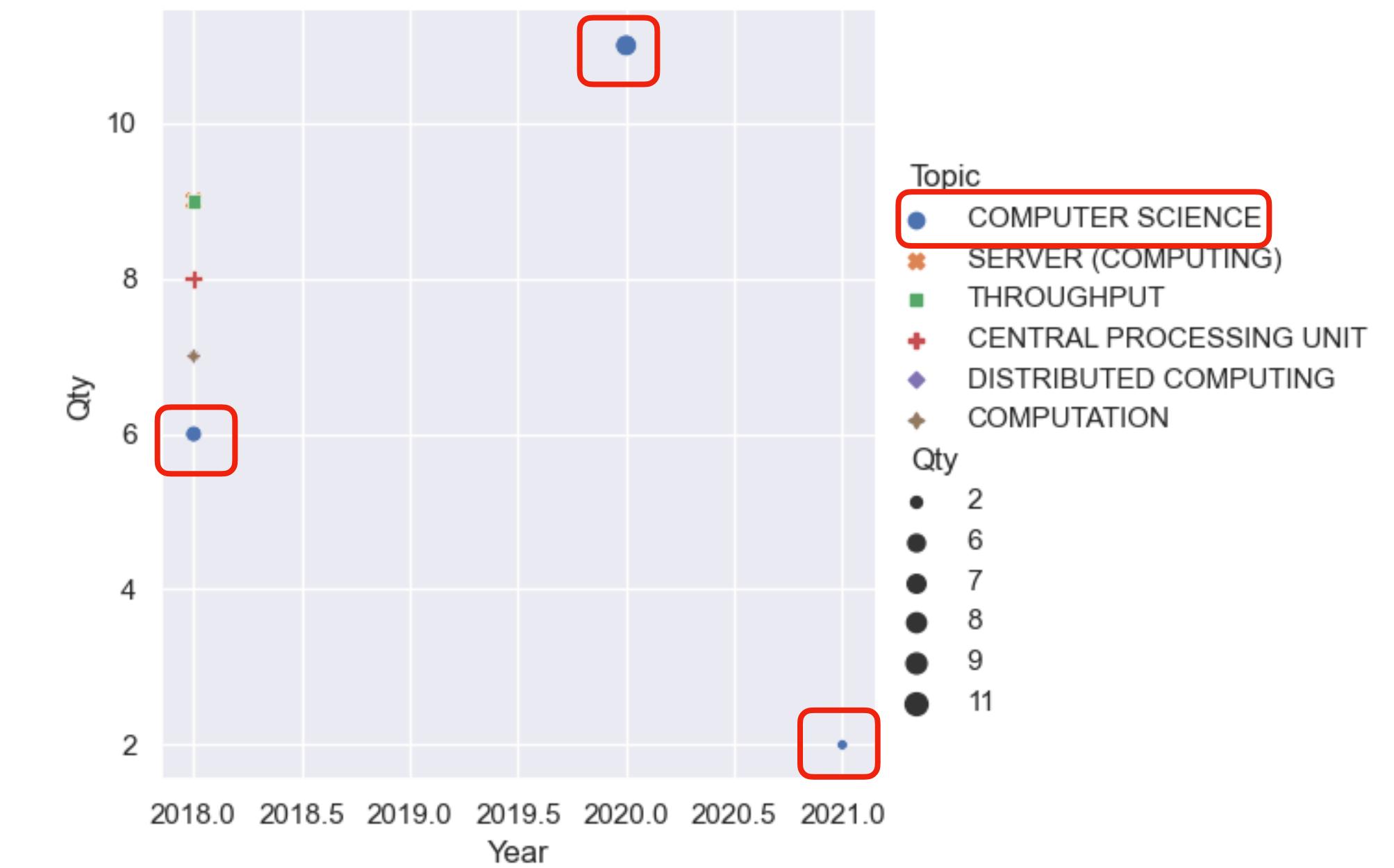
Análises e resultados



Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?



SOSP

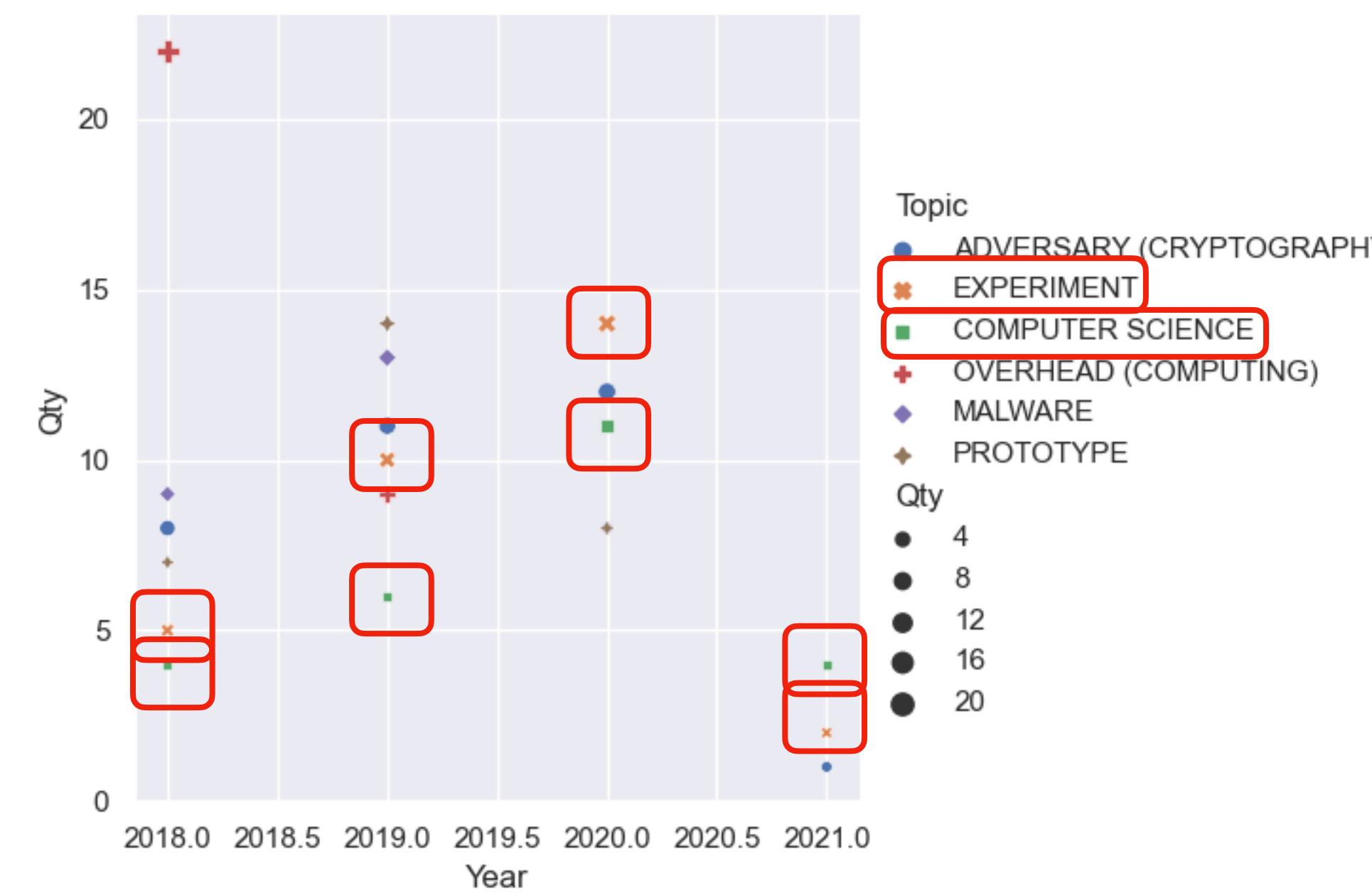


OSDI

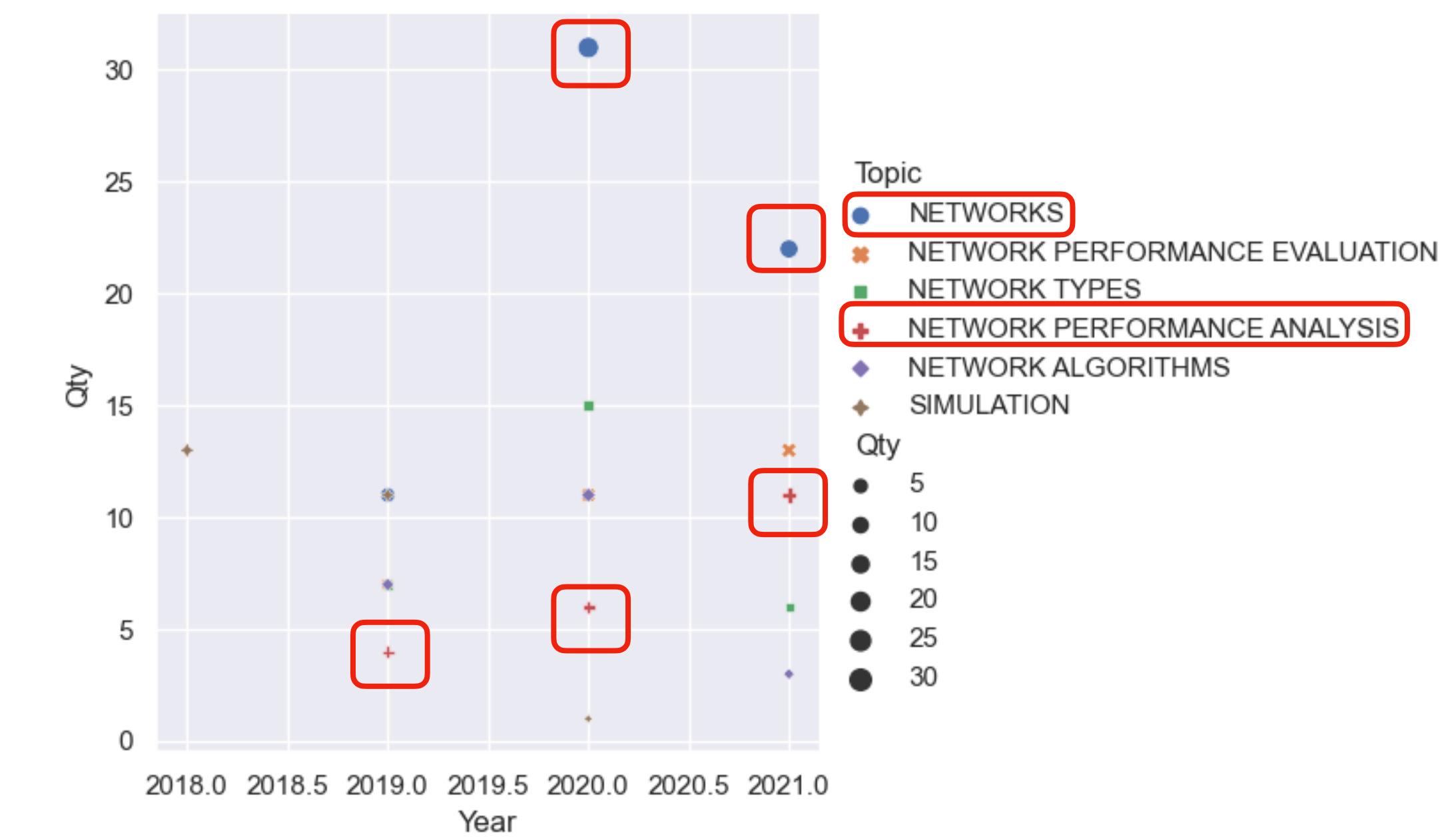
Análises e resultados



Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?



NDSS

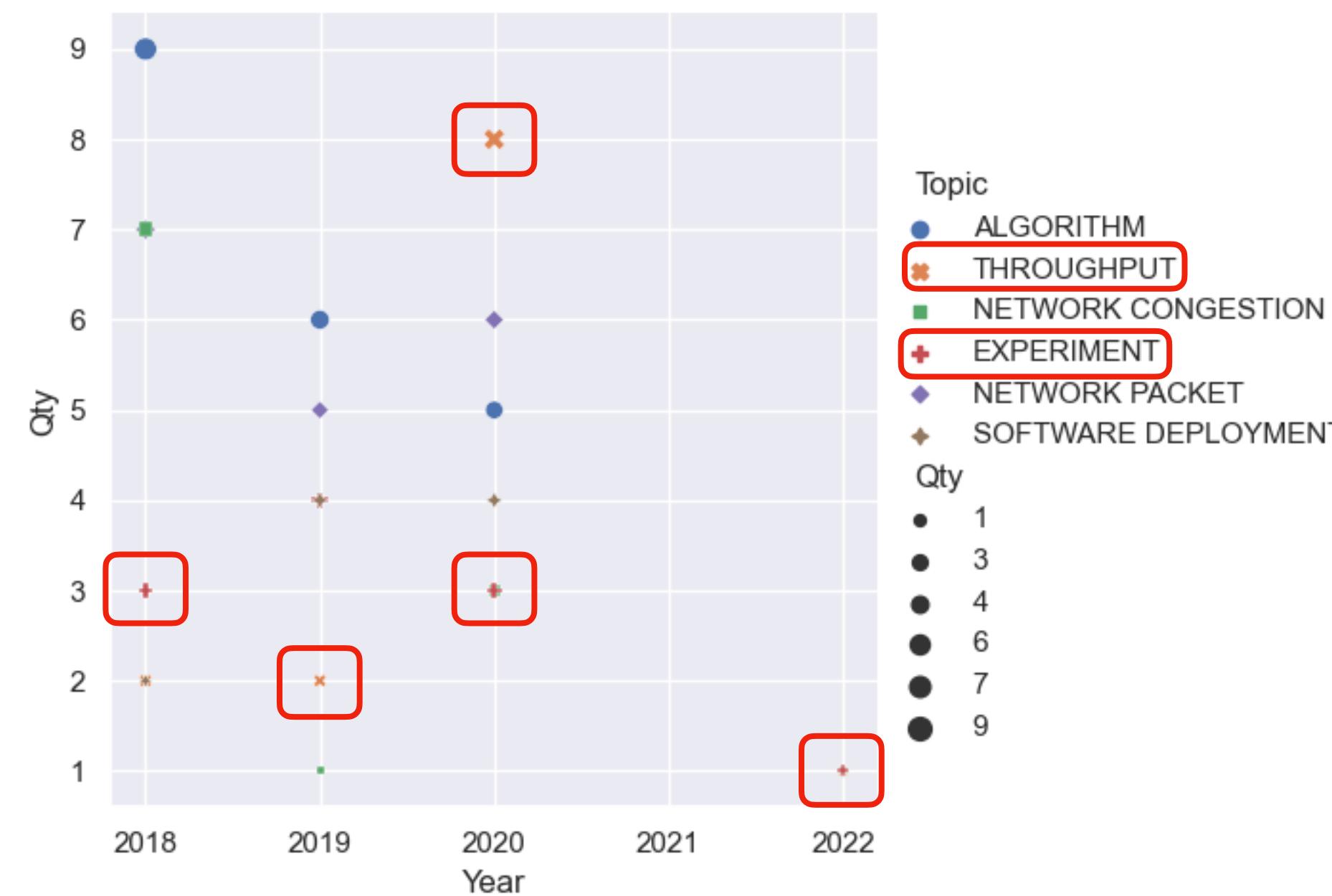


MobiHoc

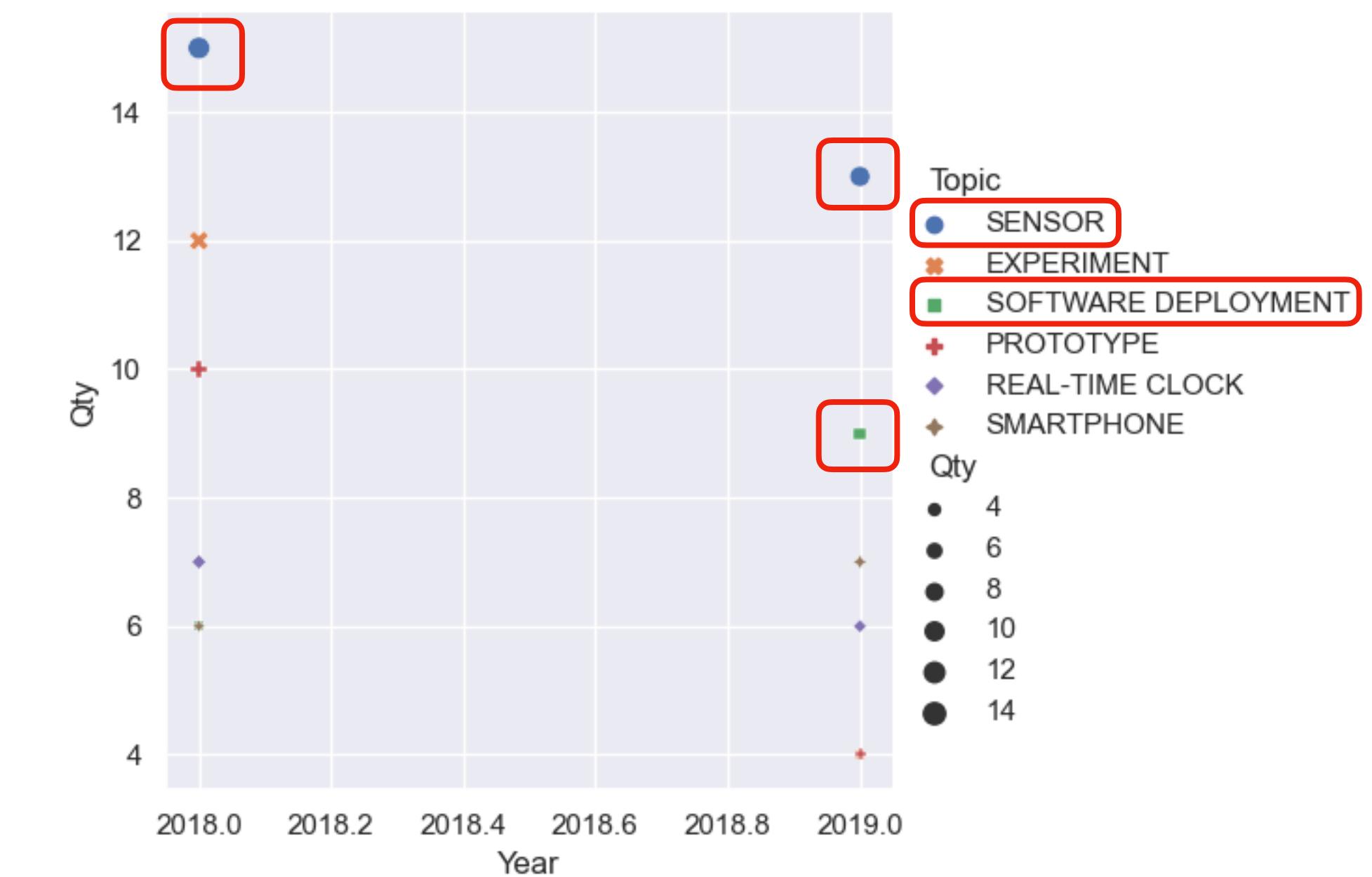
Análises e resultados



Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?



SIGCOMM

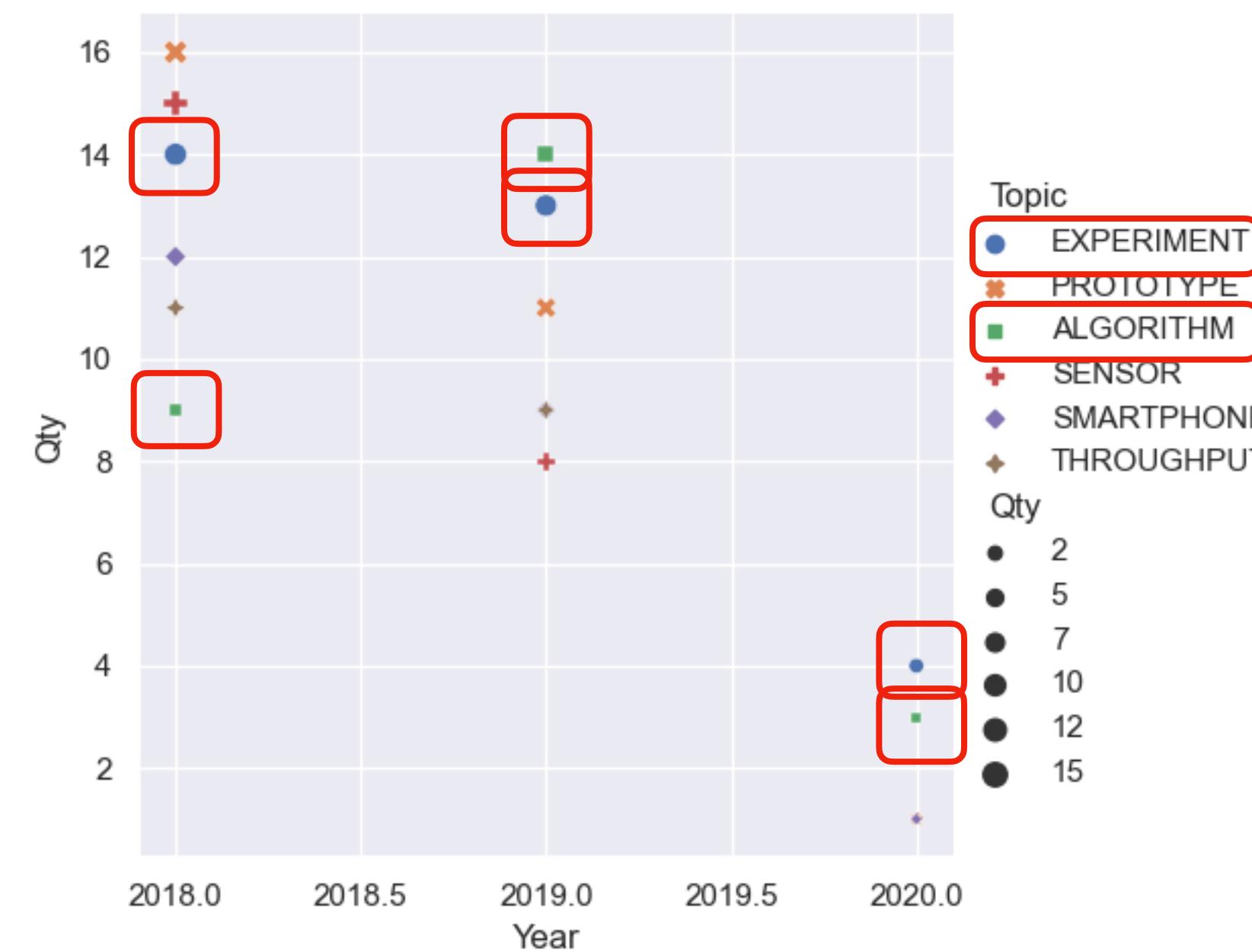


SenSys

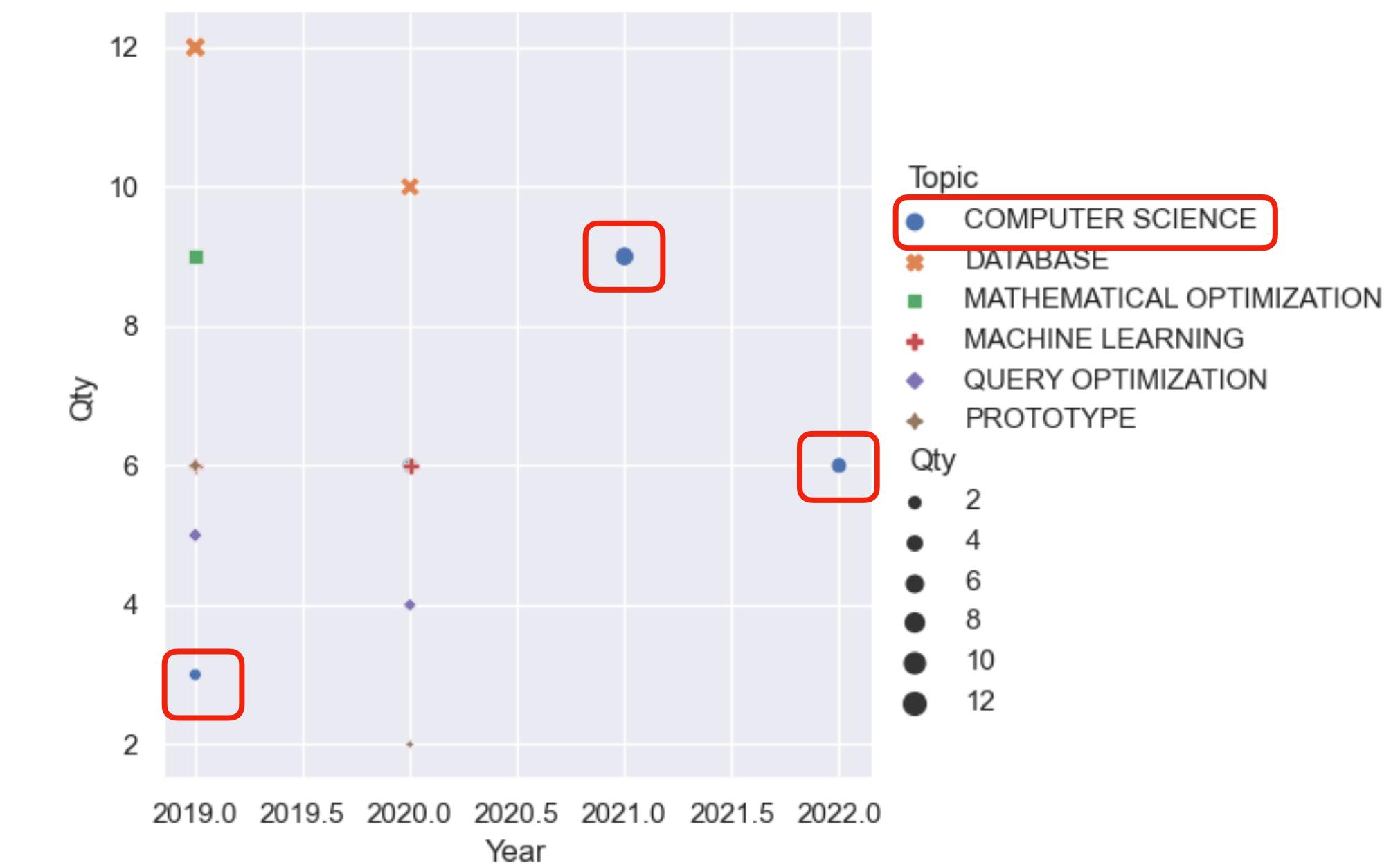
Análises e resultados



Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?



MOBICOM



CIDR

Análises e resultados



Como foram as evoluções dos tópicos publicados por cada conferência ao longo dos anos?



USENIX Security Symposium



EUROCRYPT

Problemas

Problemas



Falta de um único dataset com todos os dados necessários



Limite de requisições na API do Semantic Scholar de **100 requests por 5 minutos**

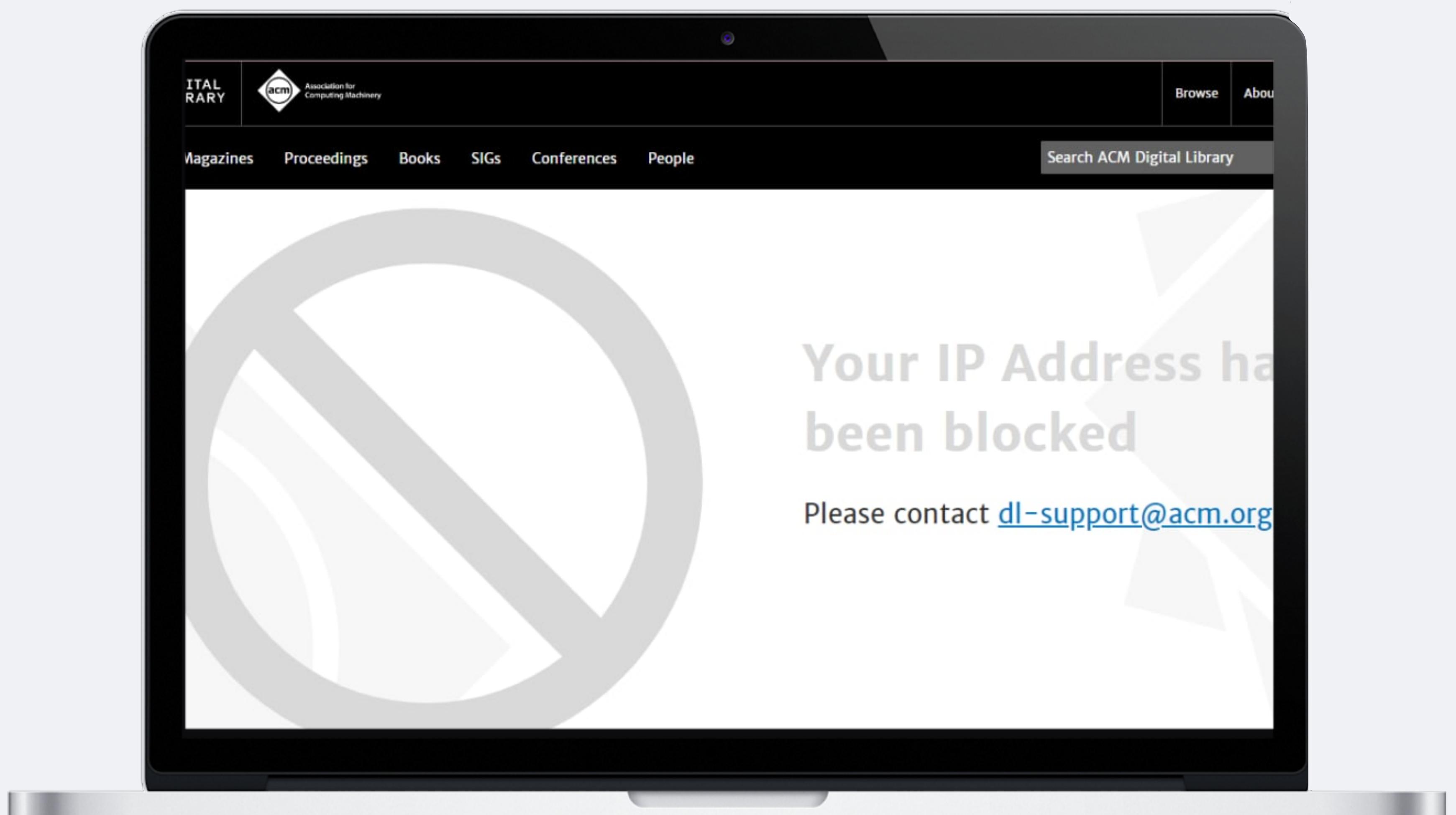


IP Bloqueado por bastante acesso ao site do DL ACM

Problemas



IP Bloqueado por bastante acesso ao site do DL ACM



Problemas



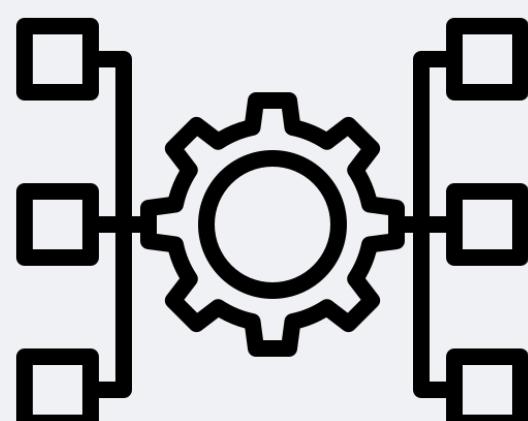
Falta de um único dataset com todos os dados necessários



Limite de requisições na API do Semantic Scholar de **100 requests por 5 minutos**



IP Bloqueado por bastante acesso ao site do DL ACM



Estruturas distintas dos artigos no site da dblp



**Trabalhos
Futuros**

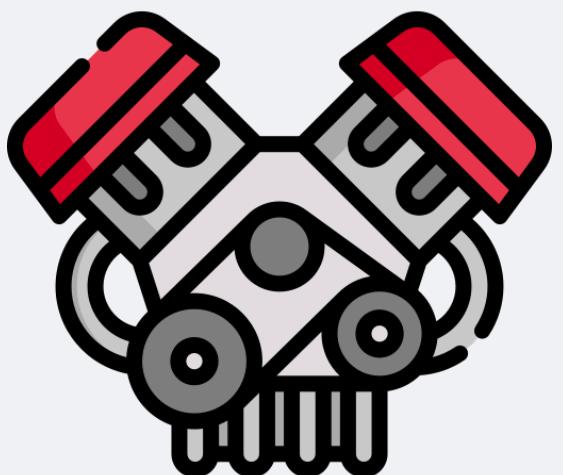
Trabalhos futuros



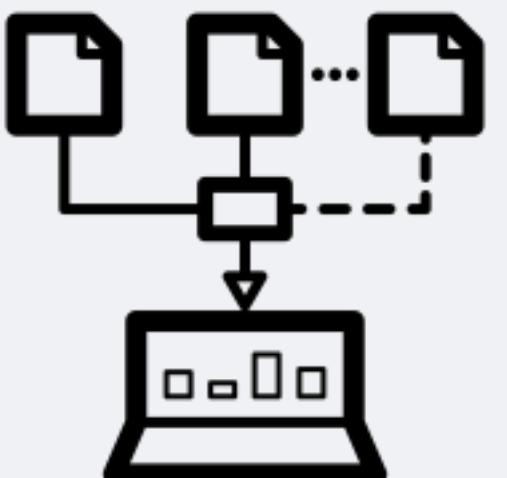
Expandir a análise para mais conferências e maiores intervalos de tempo



Expandir a análise para países dos artigos publicados nas conferências e se existe relação entre os artigos publicados em uma conferência



Implementar o sistema de recomendação que utilize a base de dados criada



Encontrar forma de trazer dados completos de todos os artigos de forma automática (uma fonte única)

Referências:

- Cuong, D. V., Nguyen, D. H., Huynh, S., Huynh, P., Gurrin, C., Dao, M.-S., Dang-Nguyen, D.-T., e Nguyen, B. T. (2020). A framework for paper submission recommendation system. In Proceedings of the 2020 International Conference on Multimedia Retrieval, ICMR '20, page 393–396, New York, NY, USA. Association for Computing Machinery.
- GOLDBERG, D. e. a. (1992). Using collaborative filtering to weave an information tapestry. Communications of the ACM, 35:61–70.
- Shao, B., L. X. . B. G. (2021). A survey of research hotspots and frontier trends of recommendation systems from the perspective of knowledge graph. Expert Systems with Applications, 165.
- Wang, D., Liang, Y., Xu, D., Feng, X., e Guan, R. (2018). A content-based recommender system for computer science publications. Knowledge-Based Systems, 157:1–9.

Fundamentos Ciência de Dados

Scientific Recommender: Análise Inicial Sobre Artigos Publicados em Conferências

