MULTIVAC PLATFORM

Large-scale scientific data at your fingertips.

Maziyar PANAHI

Chef de projet Infrastructure

ISC-PIF/CNRS

June 2023

"WORLD BEYOND DATA"

We've built scientific tools to dive into large-scale social media and Web data

Multivac is a **"Unified Big Data Analytics Platform"** launched in 2015. It hosts more than **360 billion** web pages, social media, and other digital content.

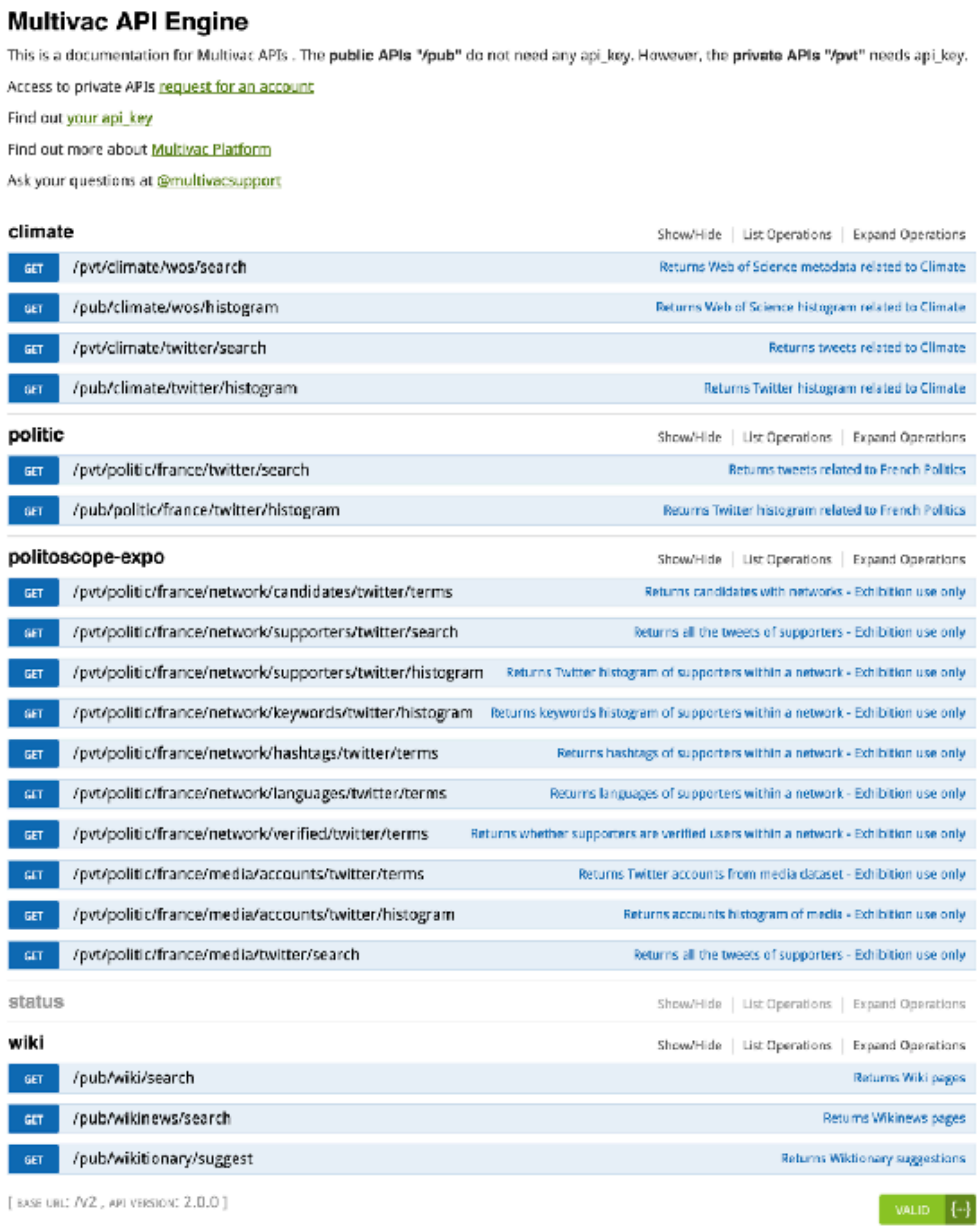
Multivac offers three main services: `Multivac Dashboards` (interactive visualizations), `Multivac API Engine` (real-time REST APIs), and `Multivac DSL` (interactive and collaborative notebooks for Big Data).

MULTIVAC API ENGINE

Multivac Platform offers a complete set of REST APIs to communicate to its data repositories. This makes it easy to get out not only the raw data, but also the aggregated and processed results.

Multivac uses Swagger to design, build and document its RESTful APIs. Swagger is a powerful open source framework backed by a large ecosystem of tools. It also follows the Open API Initiative (OAI) to standardising on how Multivac REST APIs are described.

Take a look at the demo on the right and see how you can integrate Multivac APIs inside your code.



Multivac API Engine

This is a documentation for Multivac APIs . The **public APIs "/pub"** do not need any api_key. However, the **private APIs "/pvt"** needs api_key.

Access to private APIs request for an account

Find out your api_key

Find out more about Multivac Platform

Ask your questions at @multivacsupport

climate

GET	/pvt/climate/wos/search	Returns Web of Science metadata related to Climate
GET	/pub/climate/wos/histogram	Returns Web of Science histogram related to Climate
GET	/pvt/climate/twitter/search	Returns tweets related to Climate
GET	/pub/climate/twitter/histogram	Returns Twitter histogram related to Climate

politic

GET	/pvt/politic/france/twitter/search	Returns tweets related to French Politics
GET	/pub/politic/france/twitter/histogram	Returns Twitter histogram related to French Politics

politoscope-expo

GET	/pvt/politic/france/network/candidates/twitter/terms	Returns candidates with networks - Exhibition use only
GET	/pvt/politic/france/network/supporters/twitter/search	Returns all the tweets of supporters - Exhibition use only
GET	/pvt/politic/france/network/supporters/twitter/histogram	Returns Twitter histogram of supporters within a network - Exhibition use only
GET	/pvt/politic/france/network/keywords/twitter/histogram	Returns keywords histogram of supporters within a network - Exhibition use only
GET	/pvt/politic/france/network/hashtags/twitter/terms	Returns hashtags of supporters within a network - Exhibition use only
GET	/pvt/politic/france/network/languages/twitter/terms	Returns languages of supporters within a network - Exhibition use only
GET	/pvt/politic/france/network/verified/twitter/terms	Returns whether supporters are verified users within a network - Exhibition use only
GET	/pvt/politic/france/media/accounts/twitter/terms	Returns Twitter accounts from media dataset - Exhibition use only
GET	/pvt/politic/france/media/accounts/twitter/histogram	Returns accounts histogram of media - Exhibition use only
GET	/pvt/politic/france/media/twitter/search	Returns all the tweets of supporters - Exhibition use only

status

wiki

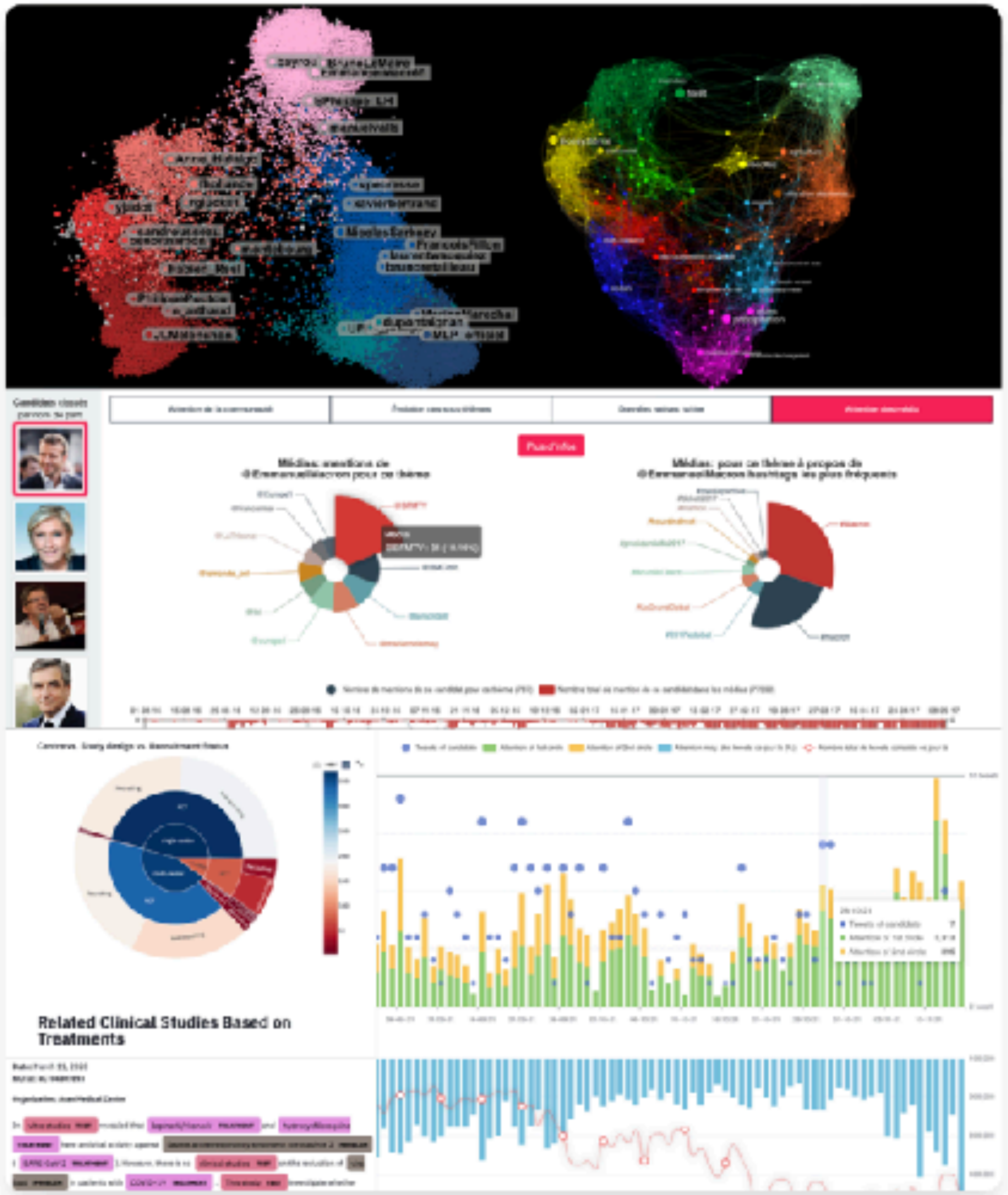
GET	/pub/wiki/search	Returns Wiki pages
GET	/pub/wikinews/search	Returns Wikinews pages
GET	/pub/wikitionary/suggest	Returns Wiktionary suggestions

[BASE URL: /V2 , API VERSION: 2.0.0]

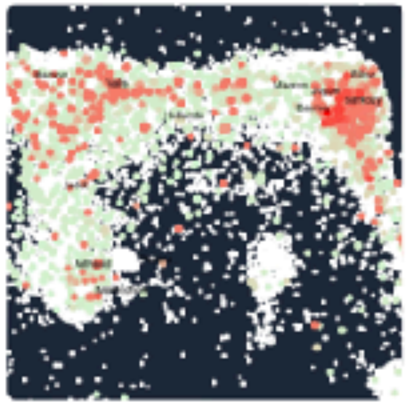
MULTIVAC DASHBOARDS

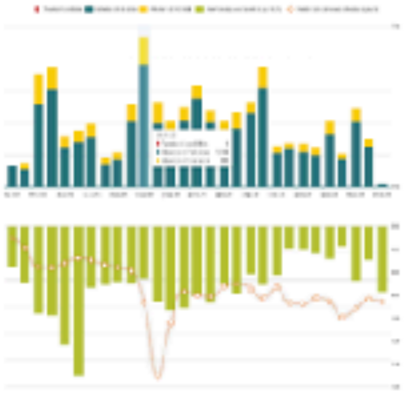
Interactive and real-time dashboards.

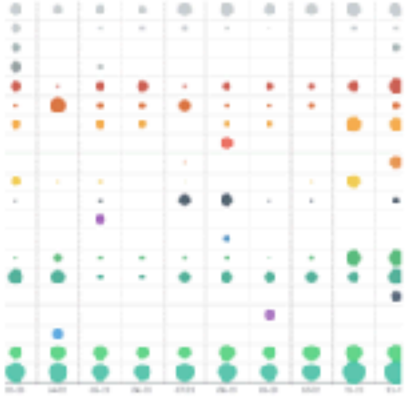
Multivac Platform offers curated dashboards to support common scientific use-cases among a variety of research topics such as Climate Change, international and French media, French political communities, Healthcare, and many more.



Journalist Dashboards


Diffusion
Political Communities


Pulse
Political Histograms


Pulse
Political Terms

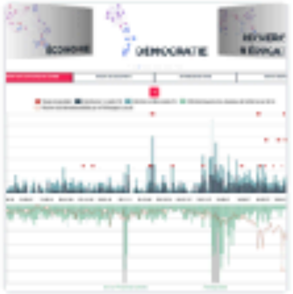
OPEN ACCESS

Public Dashboards.

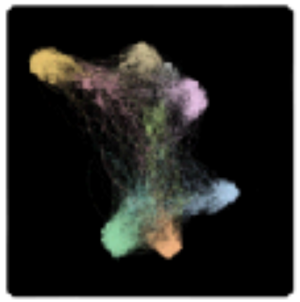
The following dashboards and interactive visualizations are open to public

Presidentielle2022

COVID-19 Dashboard


Presidentielle2017

CLIMATOSCOPE

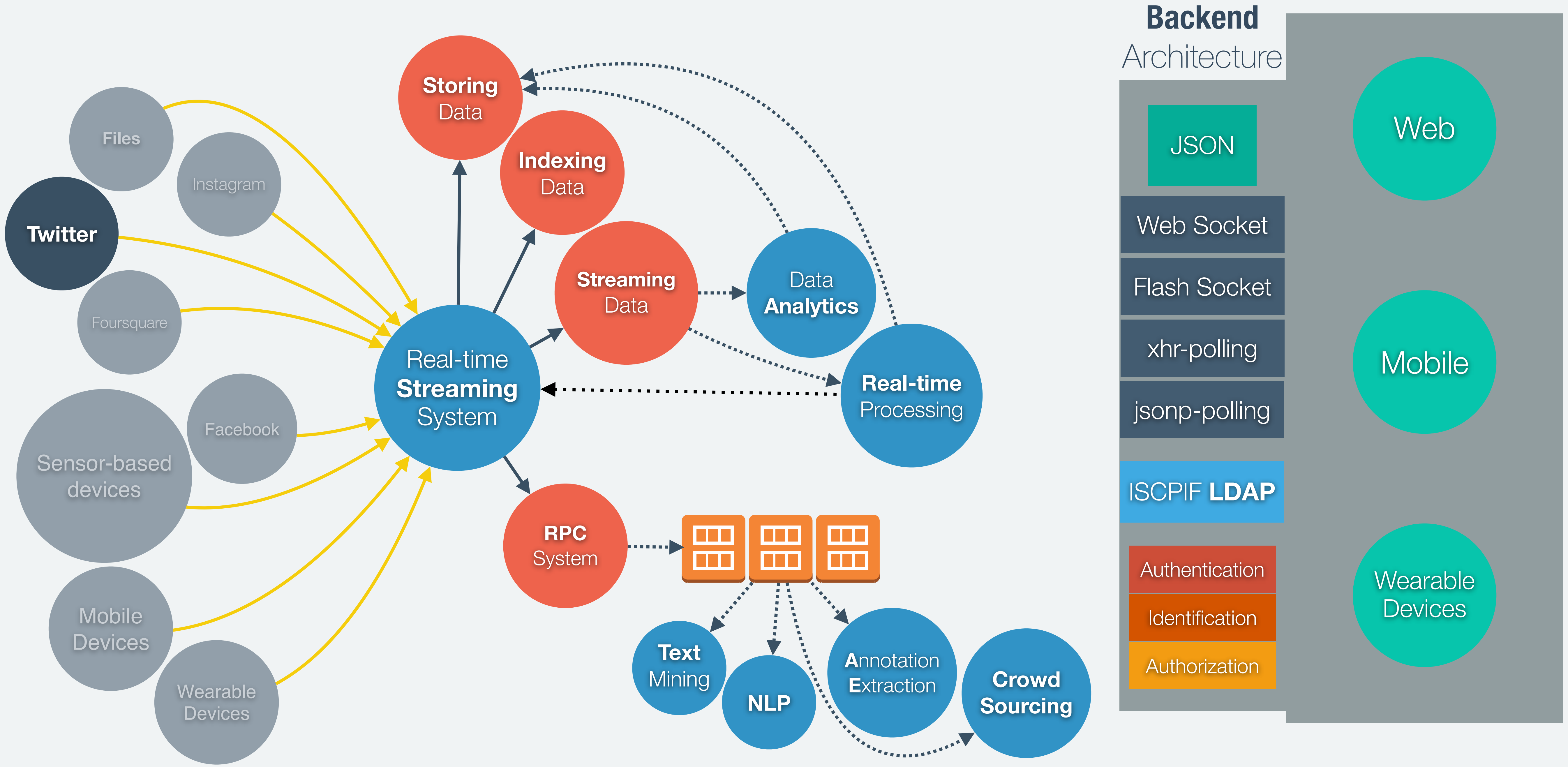

RISK RESEARCH

Real-Time Data Stream Processing

Flux de données

Système de streaming temps-réel

Utilisateur final



Backend Architecture

Files

Instagram

Twitter

Foursquare

Facebook

Sensor-based devices

Mobile Devices

Wearable Devices

Real-time **Streaming** System

Storing Data

Indexing Data

Streaming Data

Data **Analytics**

Real-time Processing

RPC System

Text Mining

NLP

Annotation **E**xtraction

Crowd Sourcing

JSON

Web Socket

Flash Socket

xhr-polling

jsonp-polling

ISCPIF **LDAP**

Authentication

Identification

Authorization

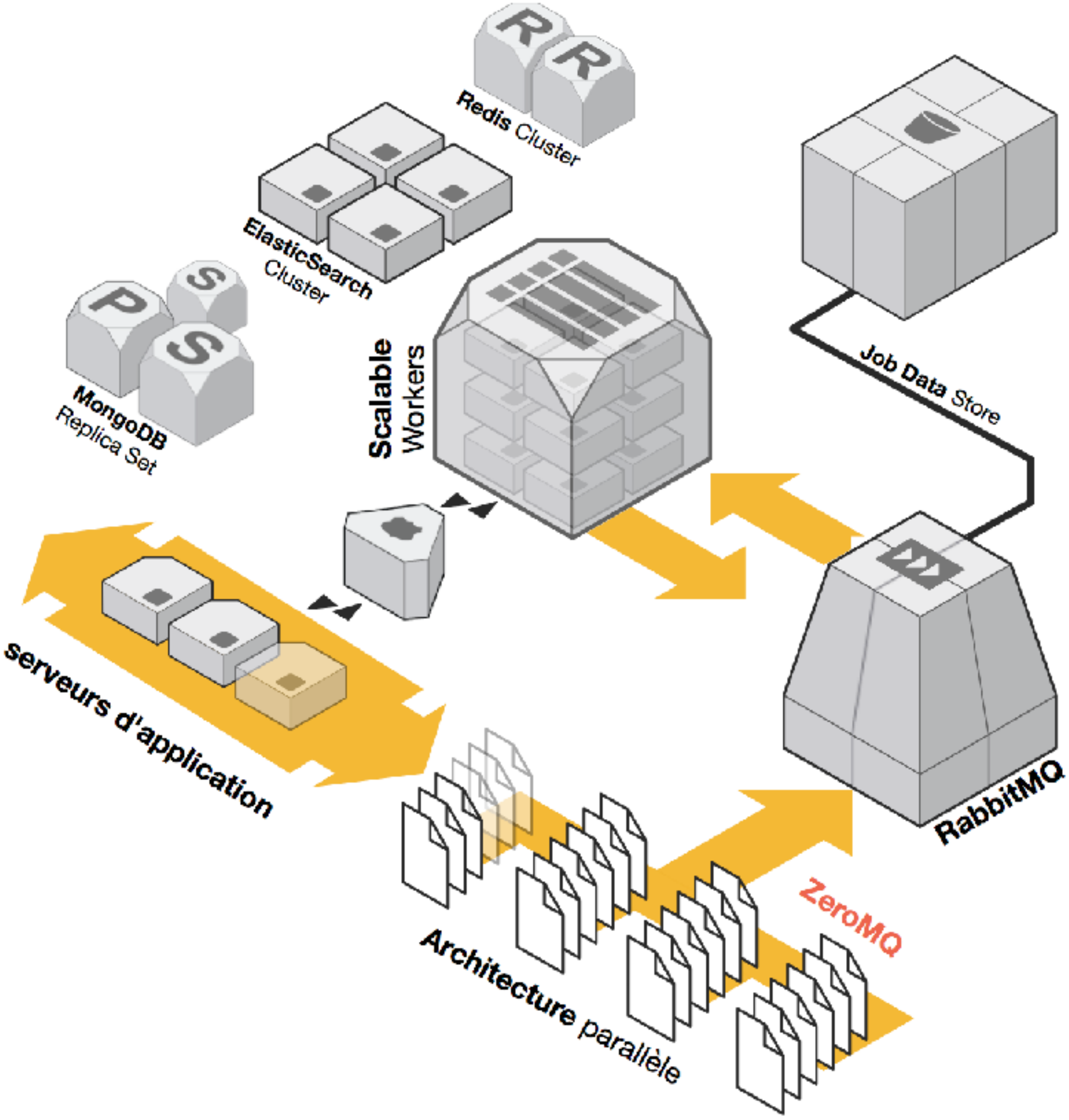
Web

Mobile

Wearable Devices

>**60 million** de données par jour
Projets: Climatique, Risk, Présidentielle 2017, etc.

High Performance Infrastructure



Redis Cluster

ElasticSearch Cluster

MongoDB Replica Set

Scalable Workers

Job Data Store

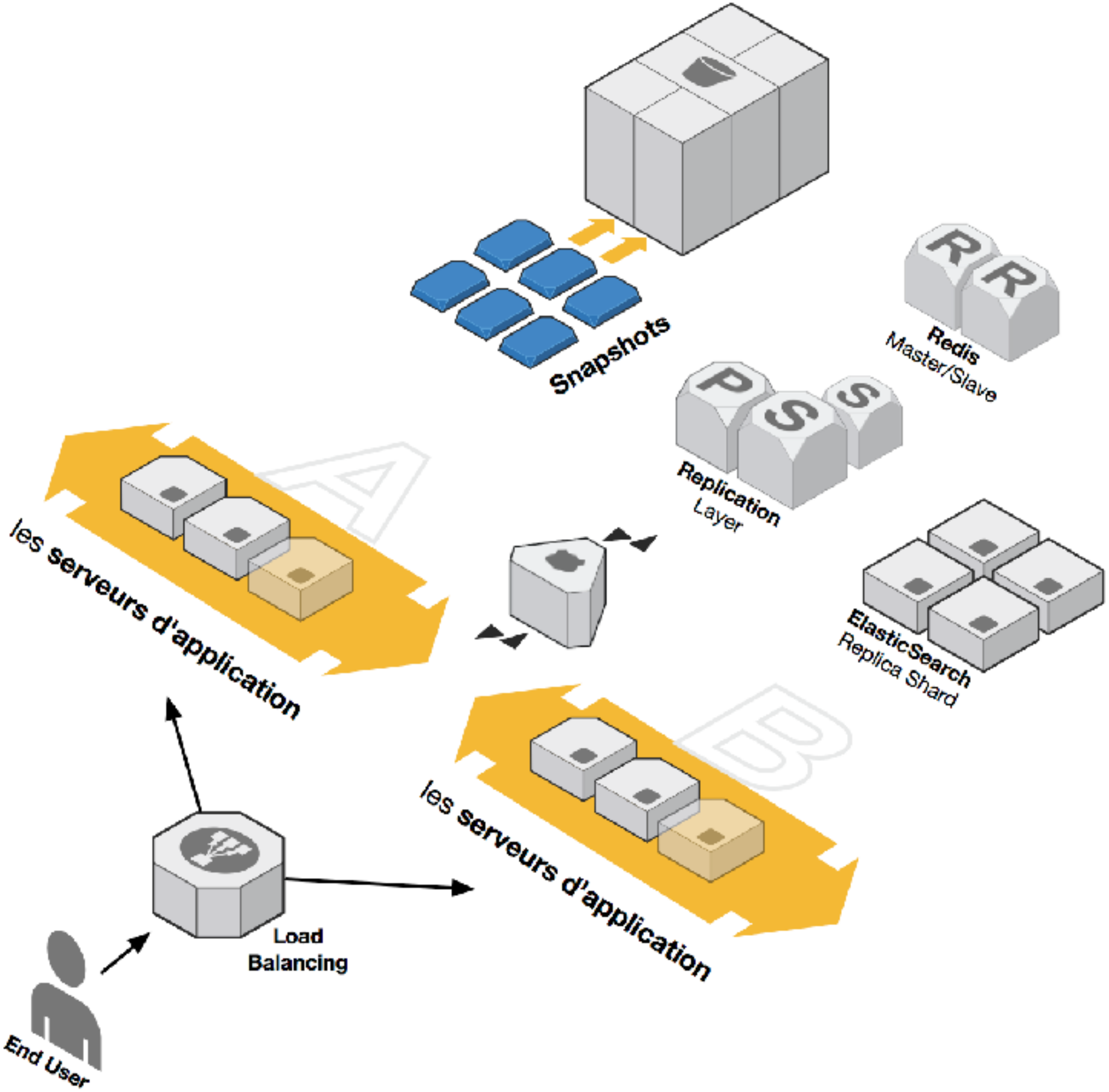
serveurs d'application

RabbitMQ

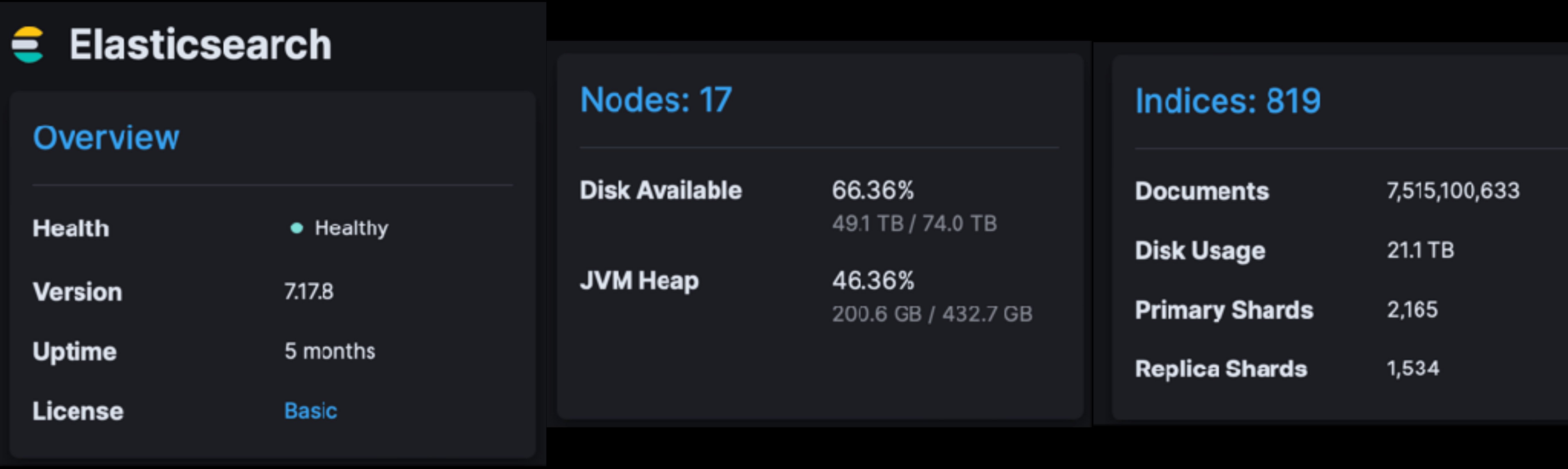
ZeroMQ

Architecture parallèle

Highly Available Infrastructure



Elasticsearch **7.17 | Multivac** Platform



Elasticsearch

Overview

Health	● Healthy
Version	7.17.8
Uptime	5 months
License	Basic

Nodes: 17

Disk Available	66.36% 49.1 TB / 74.0 TB
JVM Heap	46.36% 200.6 GB / 432.7 GB

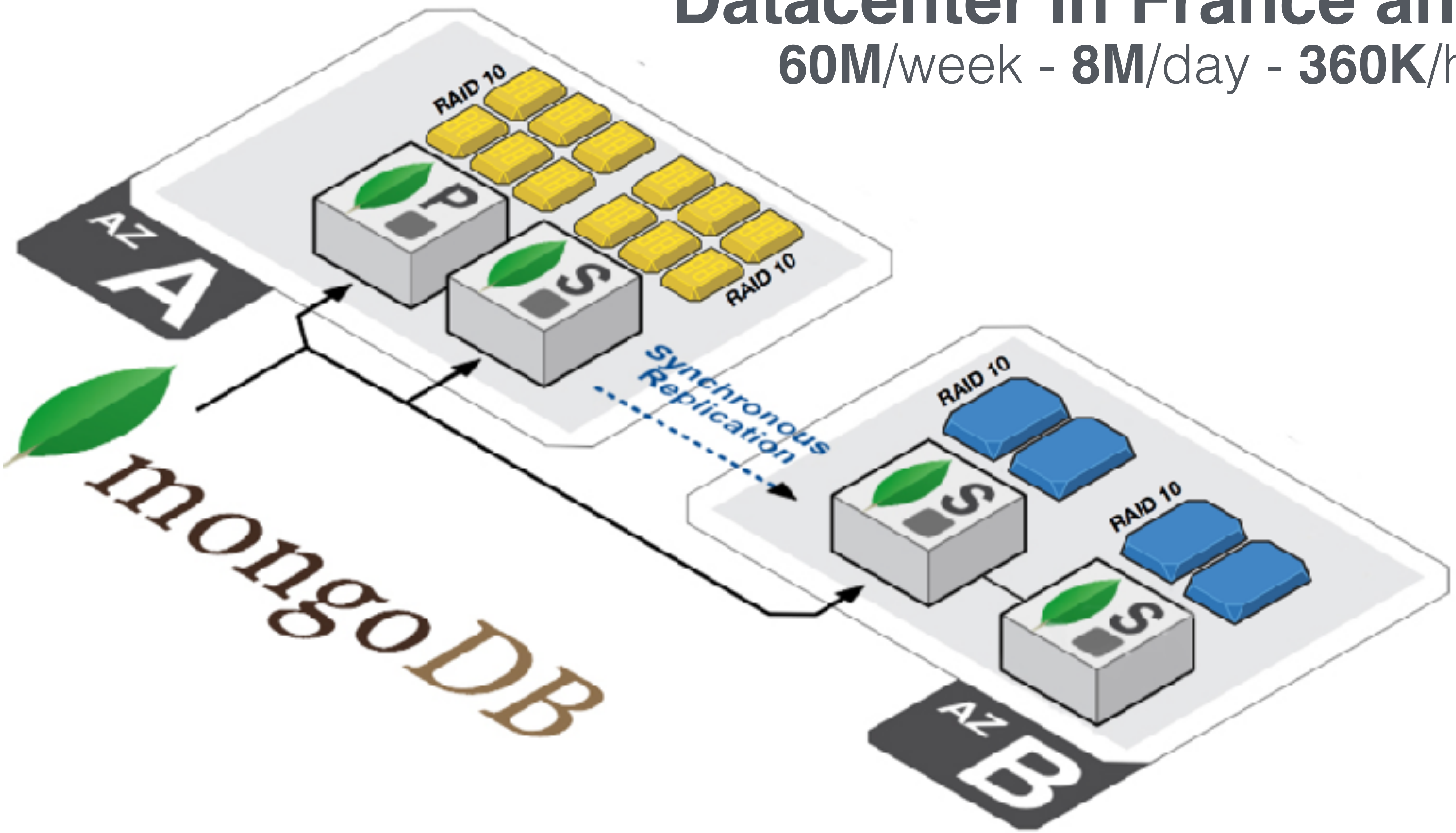
Indices: 819

Documents	7,515,100,633
Disk Usage	21.1 TB
Primary Shards	2,165
Replica Shards	1,534

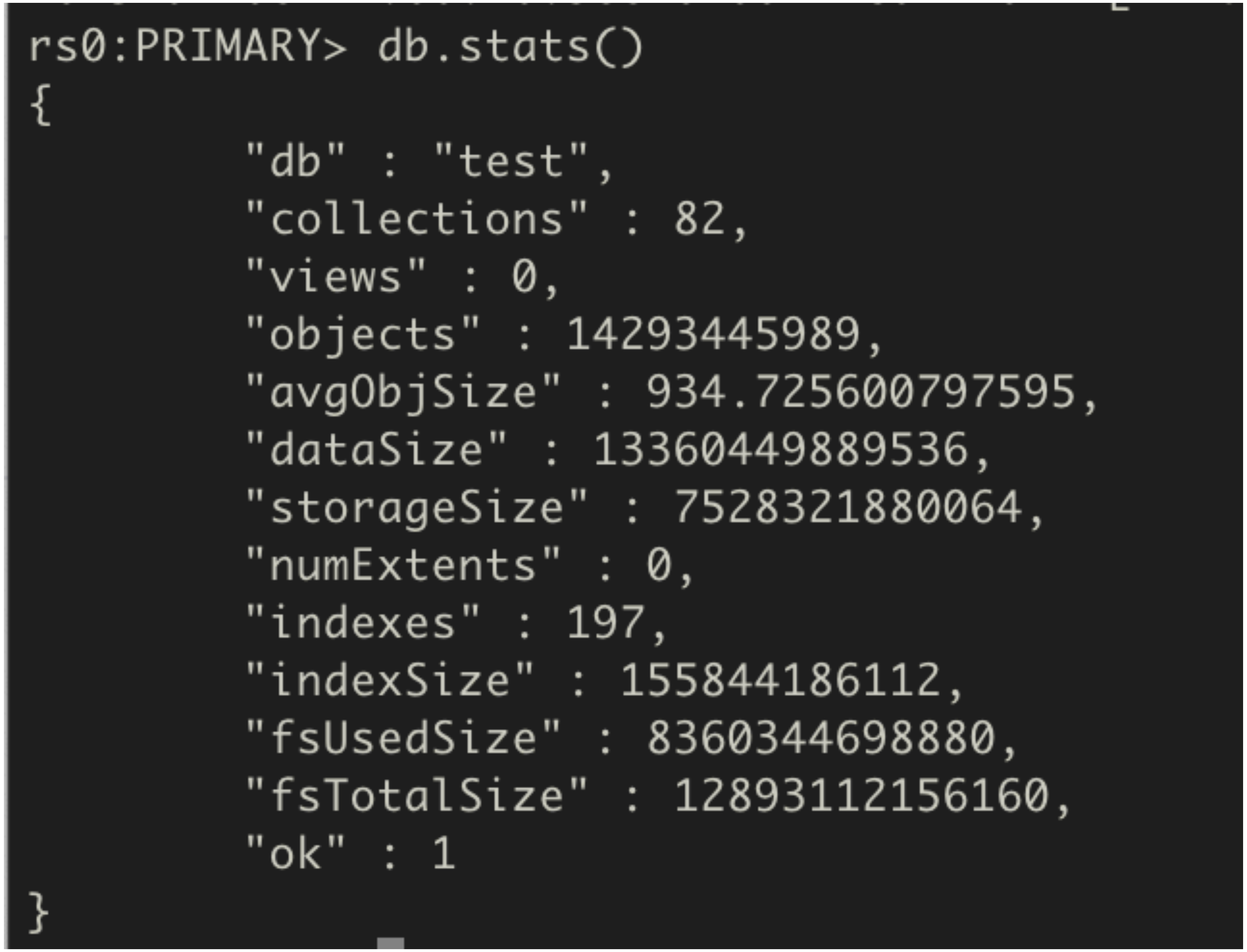
MongoDB | **Multivac** Platform

Datacenter in France and Italy

60M/week - **8M**/day - **360K**/hour

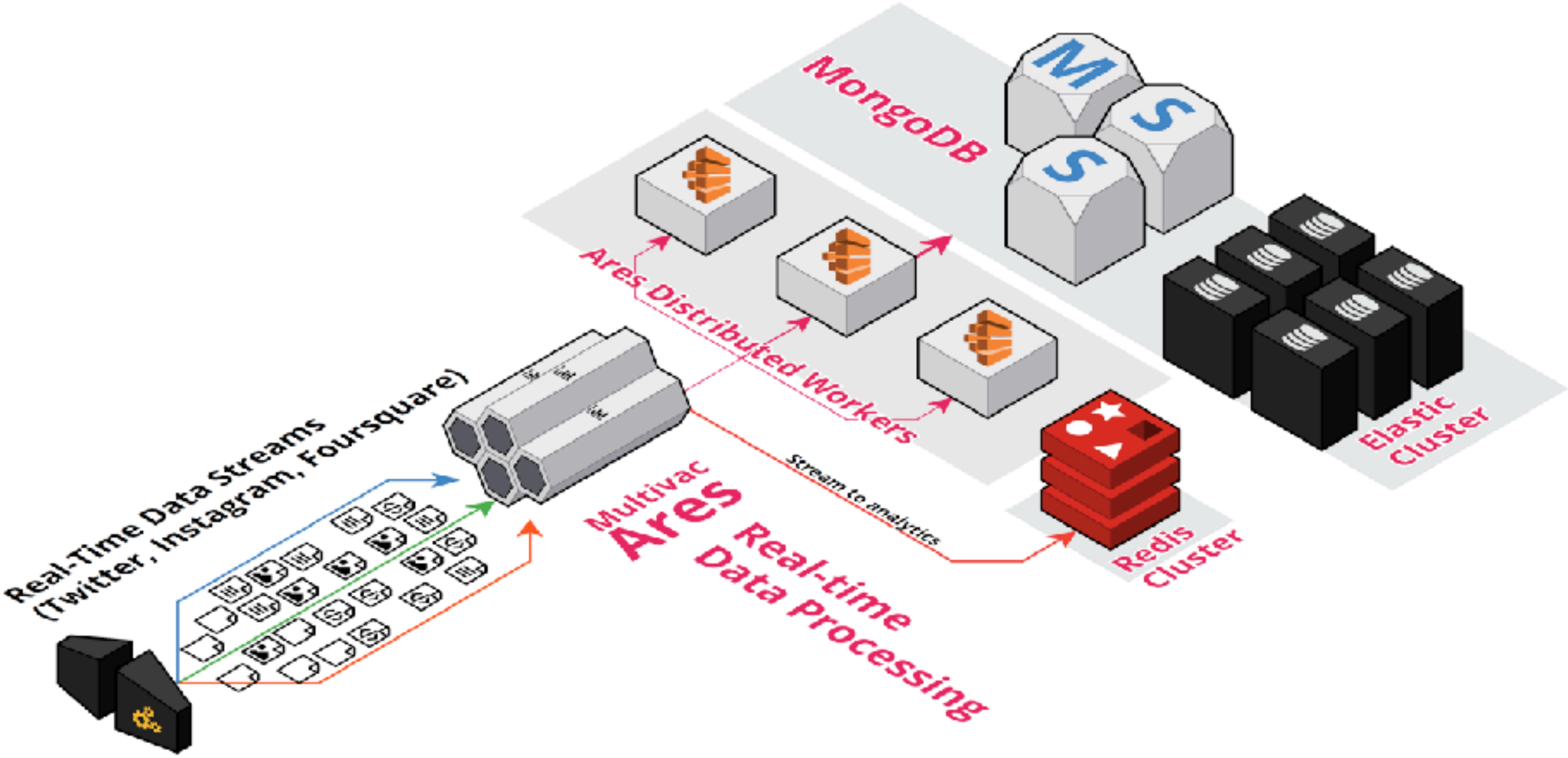


MongoDB | **Multivac** Platform



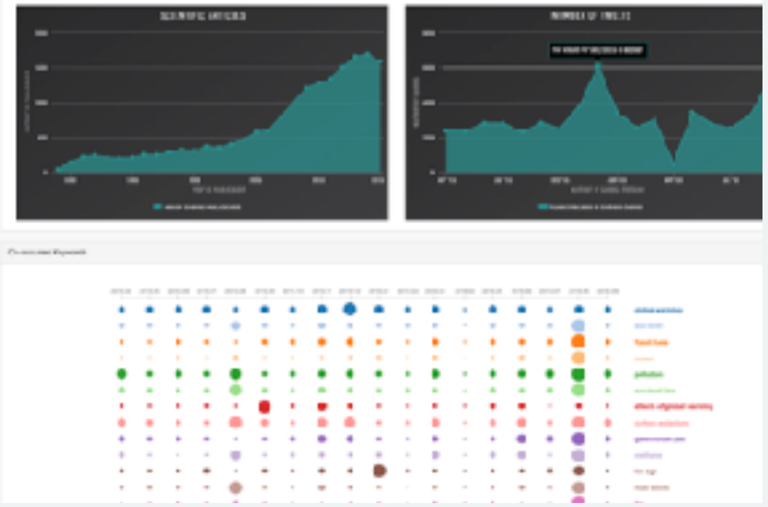
```
rs0:PRIMARY> db.stats()
{
        "db" : "test",
        "collections" : 82,
        "views" : 0,
        "objects" : 14293445989,
        "avgObjSize" : 934.725600797595,
        "dataSize" : 13360449889536,
        "storageSize" : 7528321880064,
        "numExtents" : 0,
        "indexes" : 197,
        "indexSize" : 155844186112,
        "fsUsedSize" : 8360344698880,
        "fsTotalSize" : 12893112156160,
        "ok" : 1
}
```

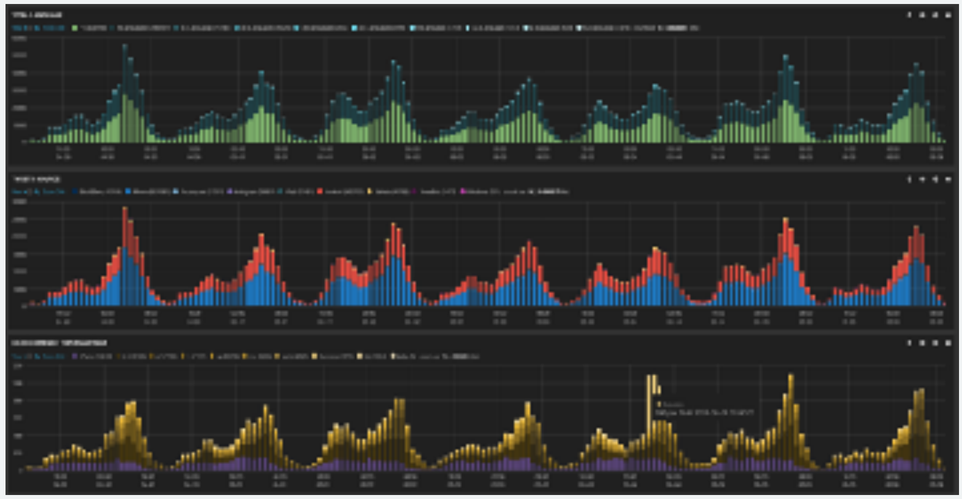
Big Data Infrastructure: Real-time streaming & computation

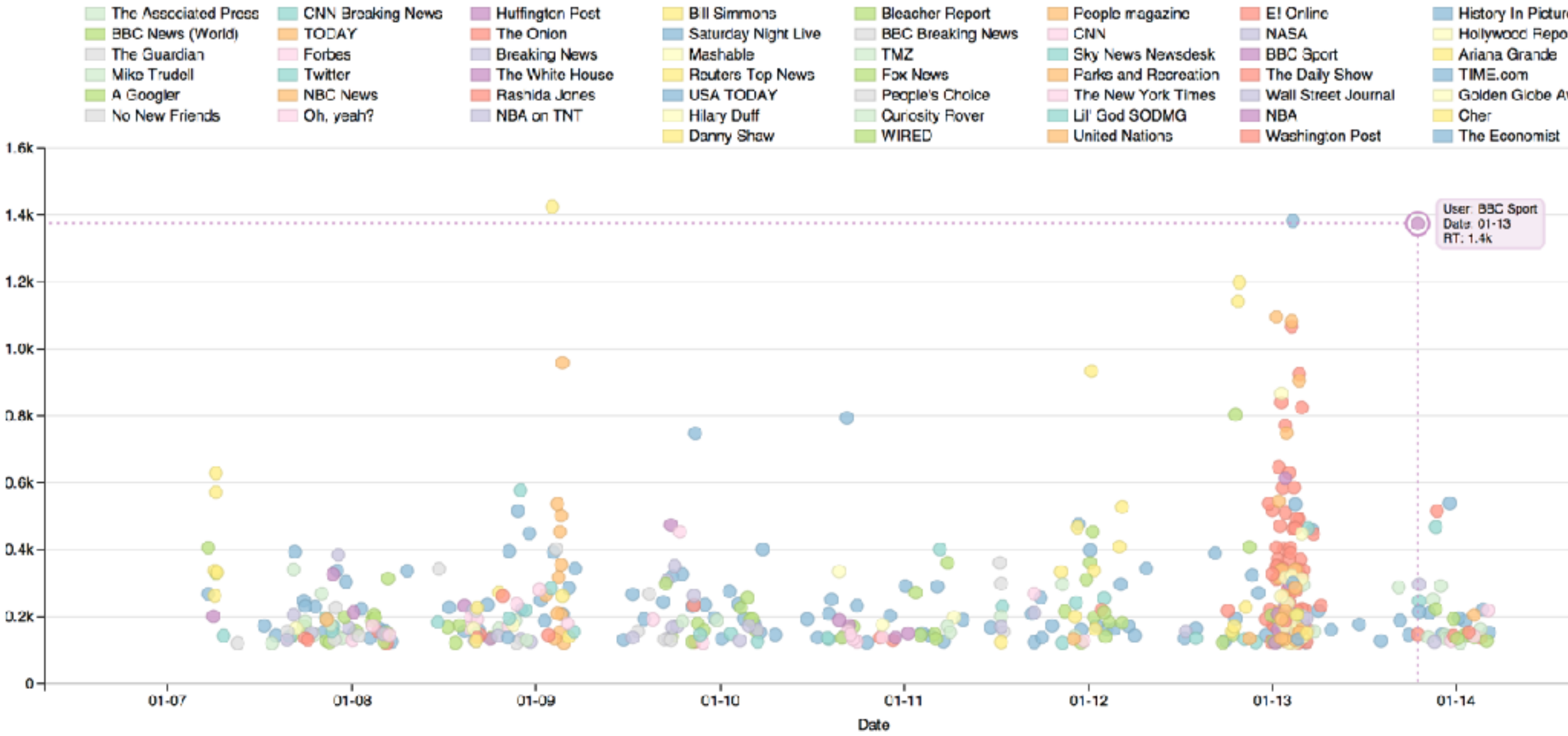


BUILT BY **MULTIVAC** PLATFORM

Current **Projects**



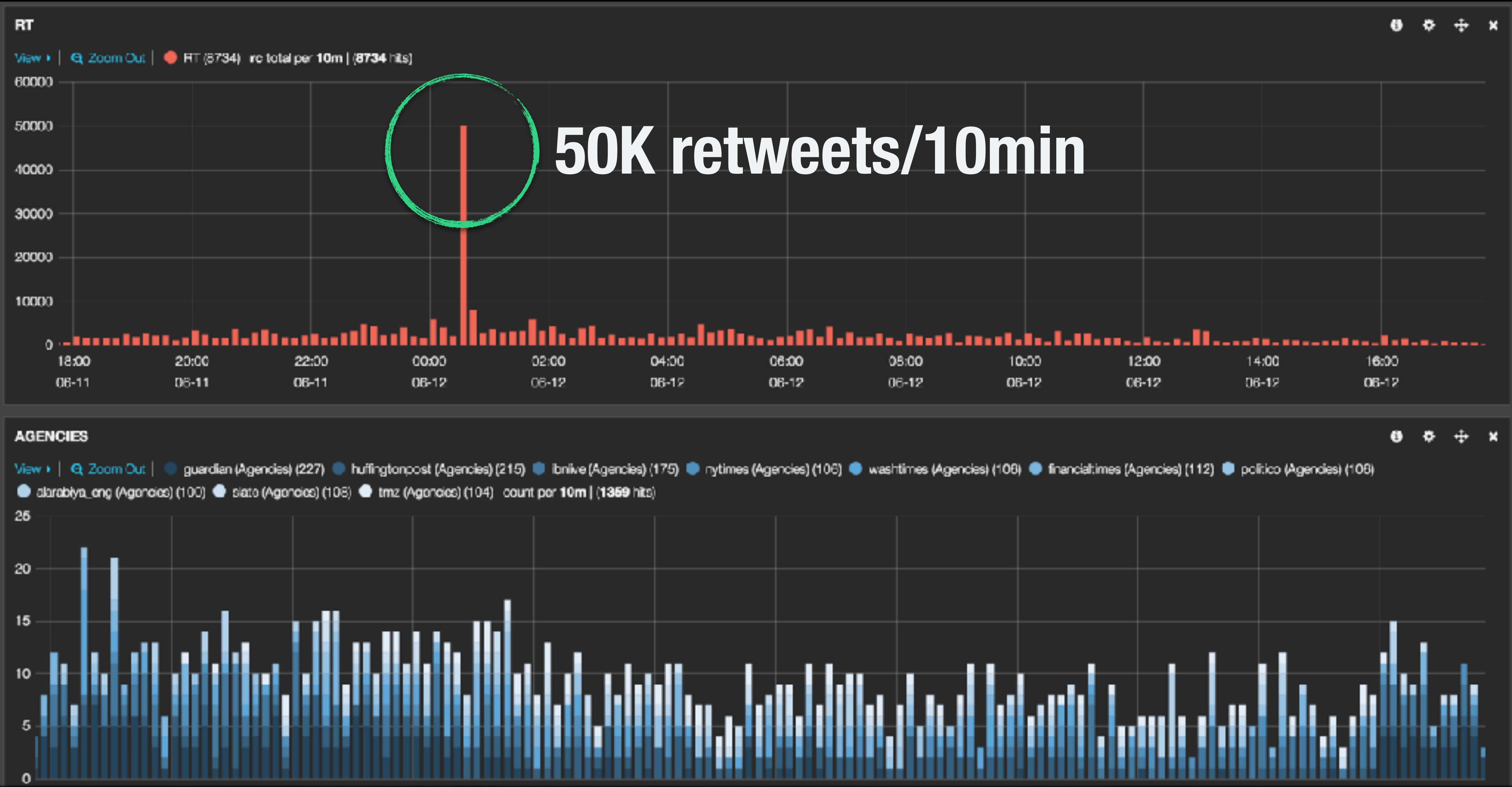




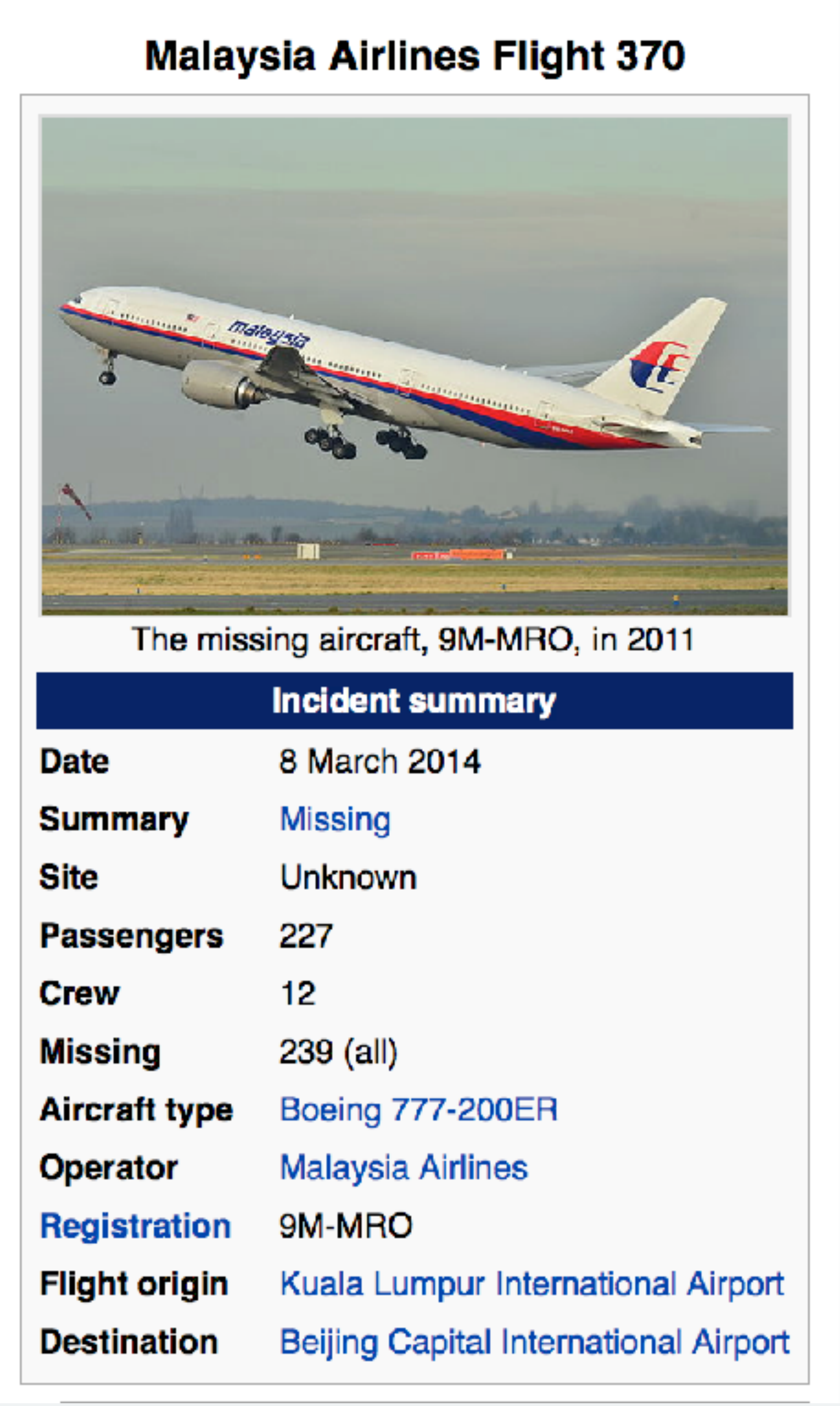
News Tracking

Real-time tracking news with highest impact of networks

Aviation Accidents


RT
View ▸ | Zoom Out | ● RT (8734) rc total per 10m | (8734 hits)
60000
50000
40000
30000
20000
10000
0
18:00 06-11
20:00 06-11
22:00 06-11
00:00 06-12
02:00 06-12
04:00 06-12
06:00 06-12
08:00 06-12
10:00 06-12
12:00 06-12
14:00 06-12
16:00 06-12
AGENCIES
View ▸ | Zoom Out | ● guardian (Agencies) (227) ● huffingtonpost (Agencies) (215) ● ibnlive (Agencies) (175) ● nytimes (Agencies) (106) ● washtimes (Agencies) (108) ● financialtimes (Agencies) (112) ● politico (Agencies) (108)
● alarabiya_eng (Agencies) (100) ● slate (Agencies) (106) ● tmz (Agencies) (104) count per 10m | (1359 hits)
25
20
15
10
5
0

50K retweets/10min

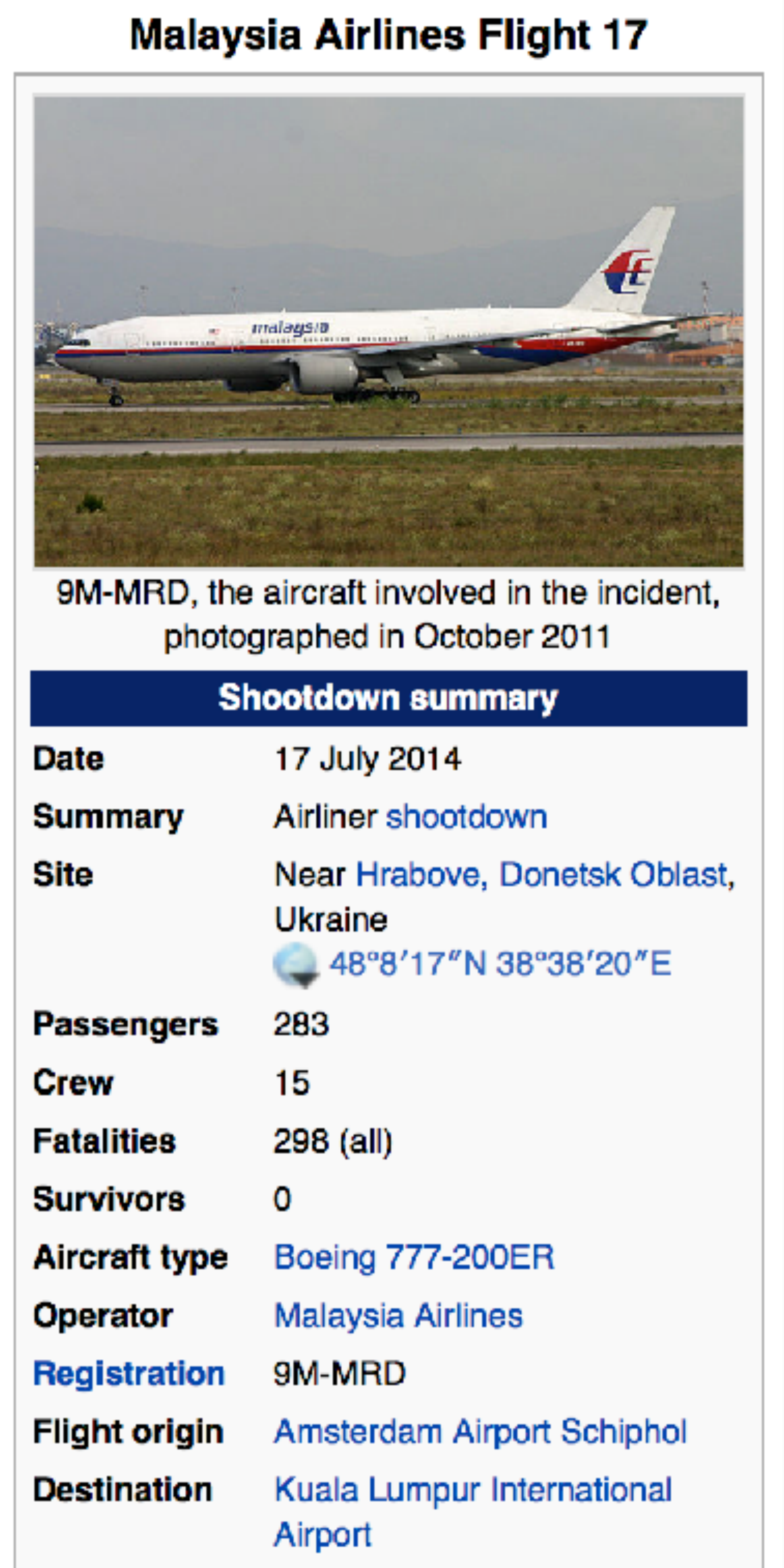


Malaysia Airlines Flight 370

The missing aircraft, 9M-MRO, in 2011

Incident summary

Date	8 March 2014
Summary	Missing
Site	Unknown
Passengers	227
Crew	12
Missing	239 (all)
Aircraft type	Boeing 777-200ER
Operator	Malaysia Airlines
Registration	9M-MRO
Flight origin	Kuala Lumpur International Airport
Destination	Beijing Capital International Airport



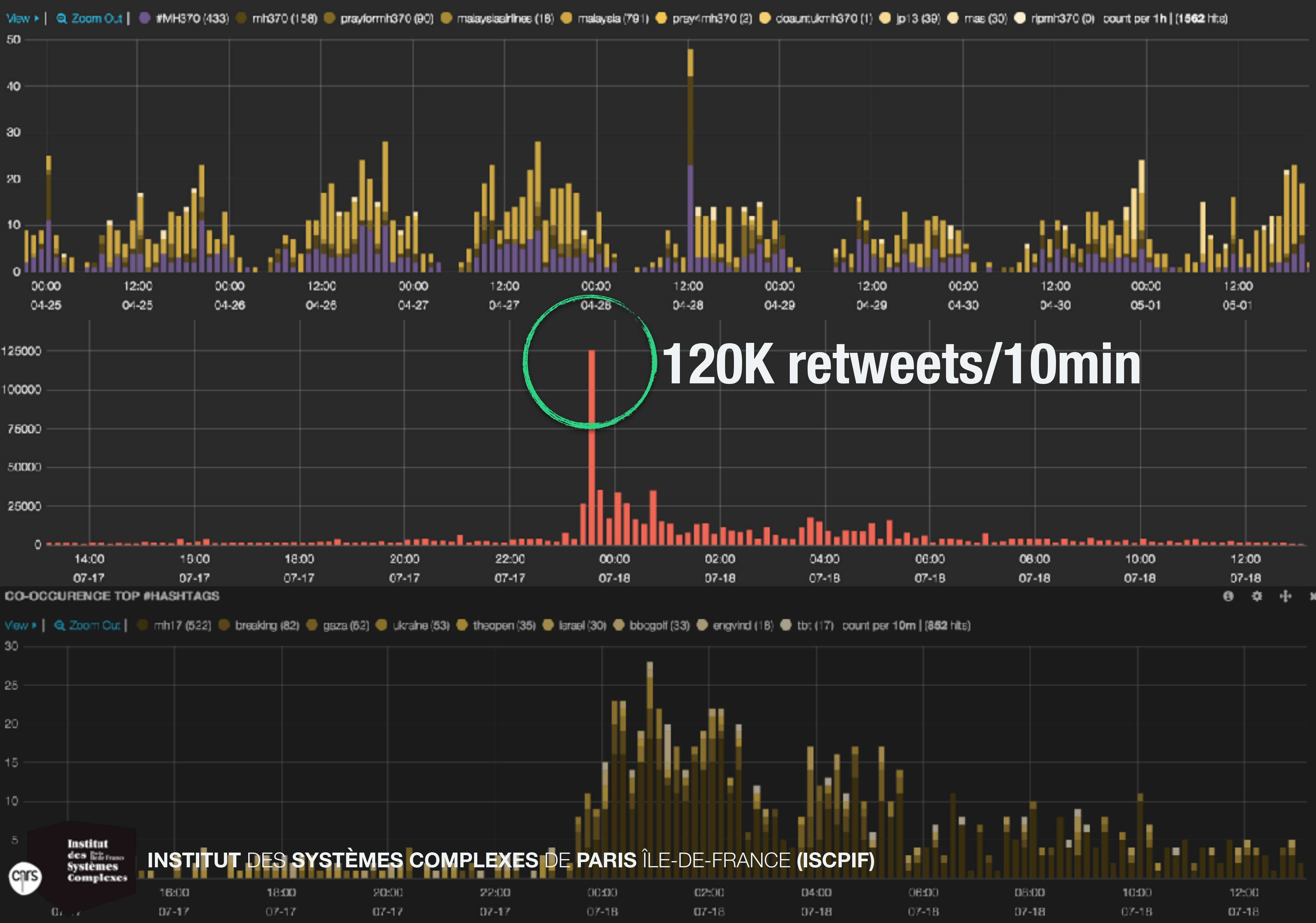
Malaysia Airlines Flight 17

9M-MRD, the aircraft involved in the incident, photographed in October 2011

Shootdown summary

Date	17 July 2014
Summary	Airliner shootdown
Site	Near Hrabove, Donetsk Oblast, Ukraine 48°8′17″N 38°38′20″E
Passengers	283
Crew	15
Fatalities	298 (all)
Survivors	0
Aircraft type	Boeing 777-200ER
Operator	Malaysia Airlines
Registration	9M-MRD
Flight origin	Amsterdam Airport Schiphol
Destination	Kuala Lumpur International Airport

Aviation Accidents



120K retweets/10min

Institut des Systèmes Complexes

INSTITUT DES **SYSTÈMES COMPLEXES** DE **PARIS** ÎLE-DE-FRANCE **(ISCPIF)**

November 24–Early December

Following the grand jury announcement, protests, some of them violent, broke out in Ferguson and other cities across the United States. Several Ferguson businesses were looted and fires set by protesters. Protests erupted in 170 cities across the U.S., including[178] St Louis, Philadelphia, Seattle, Albuquerque, New York City, Cleveland, Los Angeles, Oakland, Minneapolis, Atlanta, Chicago, and Boston.[179] Numerous media reports and legal experts criticized the process for failing to return an indictment; except in cases concerning law enforcement officers.[180][181]

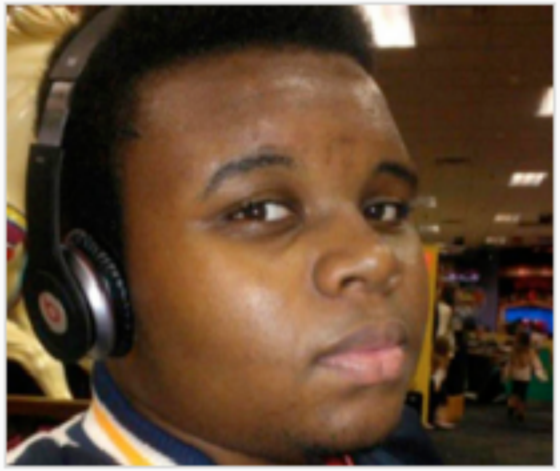
A December 2014 opinion poll that was done by Washington Post-ABC News showed that a

Missouri Highway Patrol Captain Ronald Johnson was asked to take over policing of Ferguson, as a tactical shift to reduce the violence

Police sharpshooter atop a SWAT vehicle during protests at Ferguson

Clashes between police and protesters

Michael Brown, Jr. (May 20, 1996[18] – August 9, 2014) was the son of Lesley McSpadden and Michael Brown Sr.[19] Brown graduated from Normandy High School in St. Louis County eight days before his death, completing an alternative education program.[20] At the time of his death, he was 18 years old, 6 ft 4 in (1.93 m) tall and weighed 292 lb (132 kg).[11]



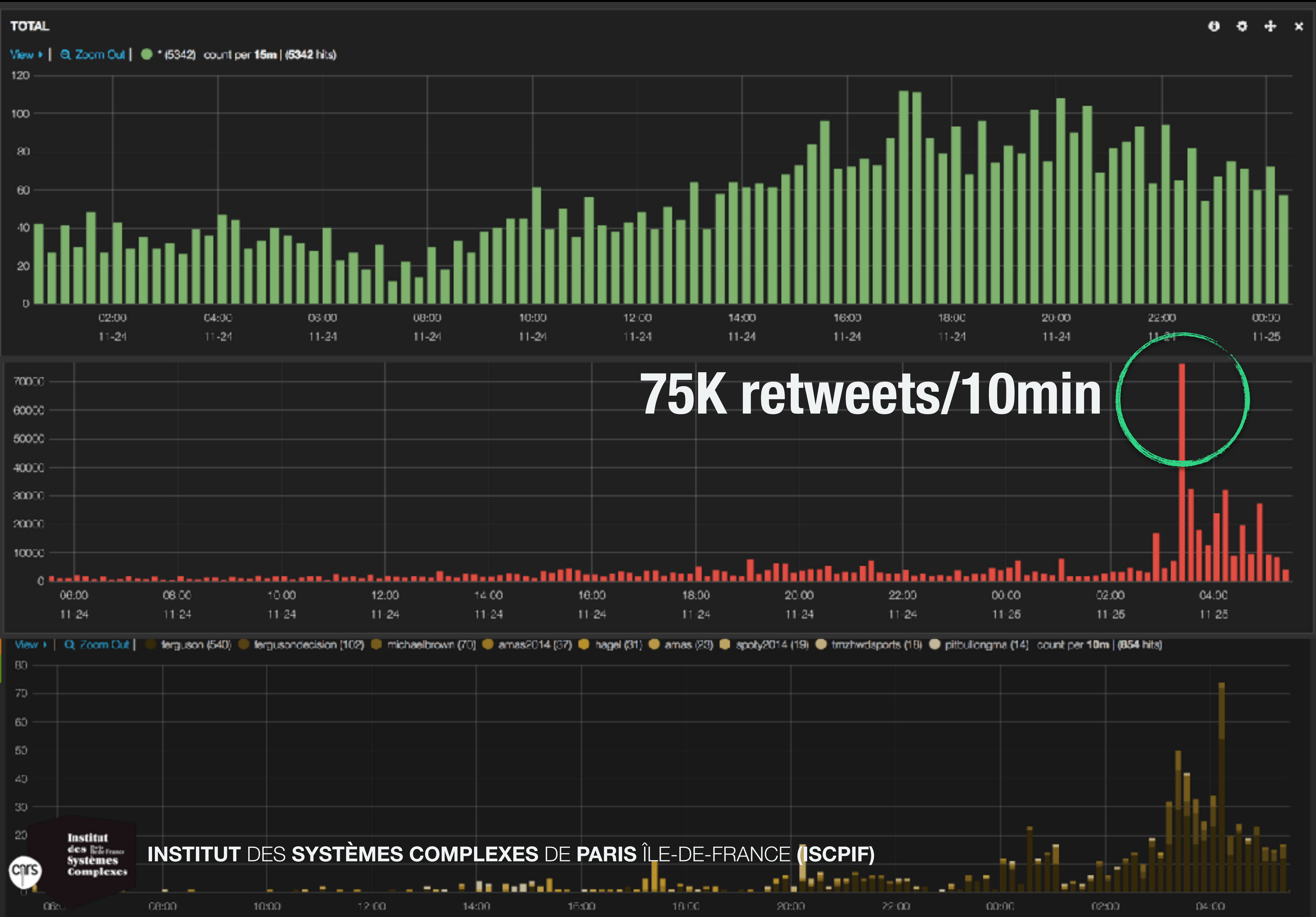
Michael Brown in a photo posted to Facebook in May 2013

Reactions to grand jury decision [edit]

The grand jury process was atypical because of significant and numerous departures from other normal grand jury proceedings. The American grand jury process operates in secret, with the proceedings, evidence and testimony rarely being released to the public in cases of no indictment. From the beginning, McCulloch desired to provide transparency to the process and had the proceeding transcribed with the intention of releasing the materials to the public if there was no indictment. [58]

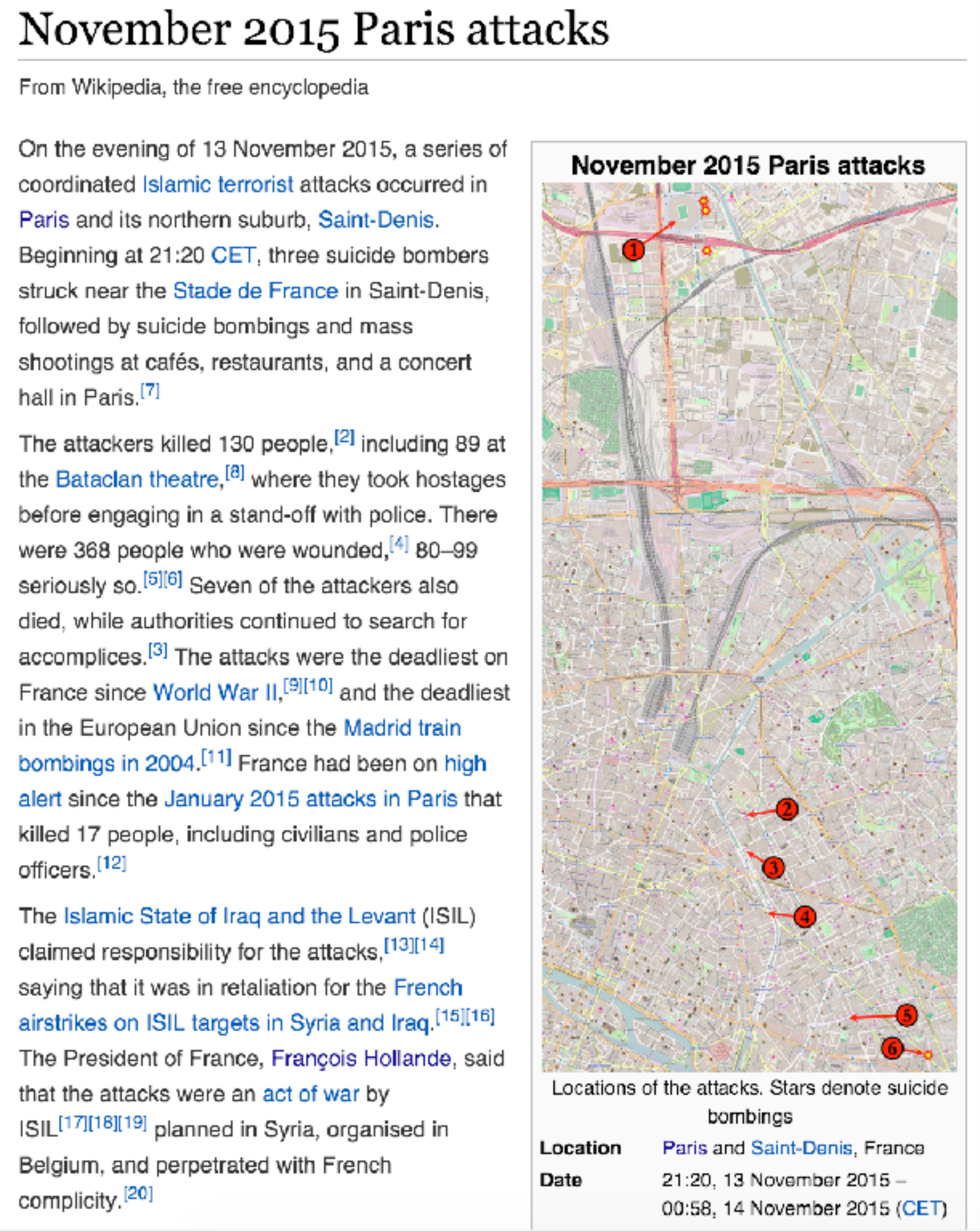
Protesters react the day following the grand jury decision in Union Square, Manhattan in New York City.

#Ferguson Michael BROWN



75K retweets/10min

INSTITUT DES **SYSTÈMES COMPLEXES** DE **PARIS** ÎLE-DE-FRANCE **(ISCPIF)**



November 2015 Paris attacks

From Wikipedia, the free encyclopedia

On the evening of 13 November 2015, a series of coordinated Islamic terrorist attacks occurred in Paris and its northern suburb, Saint-Denis. Beginning at 21:20 CET, three suicide bombers struck near the Stade de France in Saint-Denis, followed by suicide bombings and mass shootings at cafés, restaurants, and a concert hall in Paris.[7]

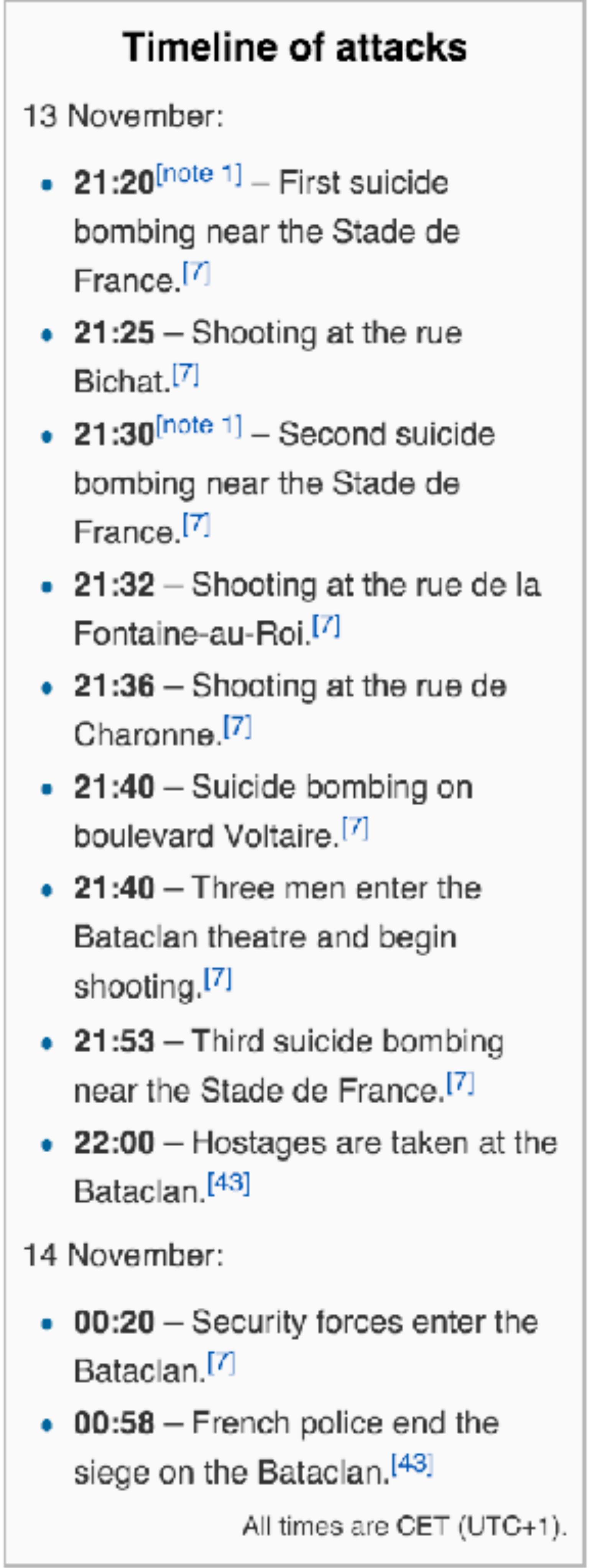
The attackers killed 130 people,[2] including 89 at the Bataclan theatre,[8] where they took hostages before engaging in a stand-off with police. There were 368 people who were wounded,[4] 80–99 seriously so.[5][6] Seven of the attackers also died, while authorities continued to search for accomplices.[3] The attacks were the deadliest on France since World War II,[9][10] and the deadliest in the European Union since the Madrid train bombings in 2004.[11] France had been on high alert since the January 2015 attacks in Paris that killed 17 people, including civilians and police officers.[12]

The Islamic State of Iraq and the Levant (ISIL) claimed responsibility for the attacks,[13][14] saying that it was in retaliation for the French airstrikes on ISIL targets in Syria and Iraq.[15][16] The President of France, François Hollande, said that the attacks were an act of war by ISIL[17][18][19] planned in Syria, organised in Belgium, and perpetrated with French complicity.[20]

November 2015 Paris attacks

Locations of the attacks. Stars denote suicide bombings

Location	Paris and Saint-Denis, France
Date	21:20, 13 November 2015 – 00:58, 14 November 2015 (CET)



Timeline of attacks

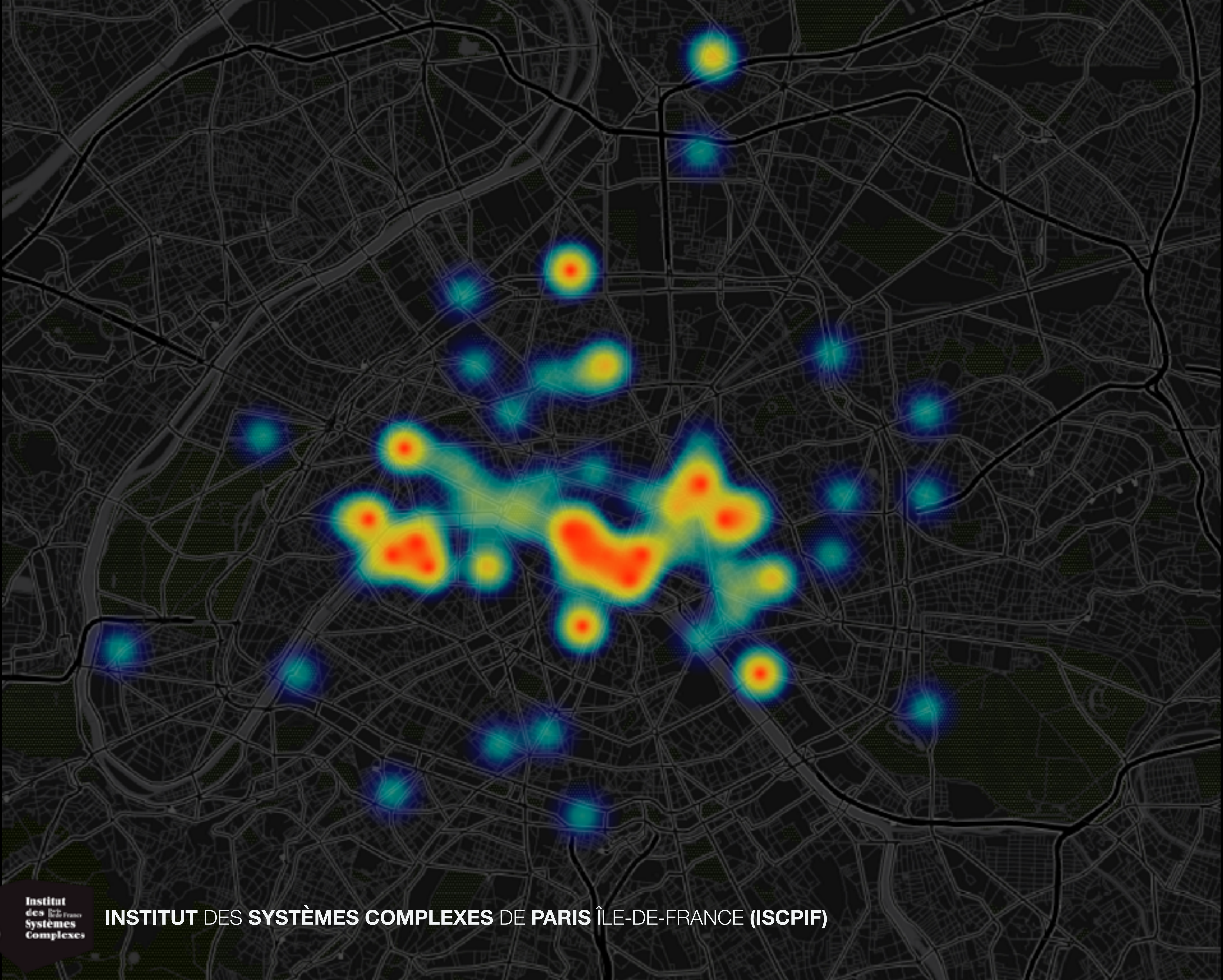
13 November:

- **21:20**[note 1] – First suicide bombing near the Stade de France.[7]
- **21:25** – Shooting at the rue Bichat.[7]
- **21:30**[note 1] – Second suicide bombing near the Stade de France.[7]
- **21:32** – Shooting at the rue de la Fontaine-au-Roi.[7]
- **21:36** – Shooting at the rue de Charonne.[7]
- **21:40** – Suicide bombing on boulevard Voltaire.[7]
- **21:40** – Three men enter the Bataclan theatre and begin shooting.[7]
- **21:53** – Third suicide bombing near the Stade de France.[7]
- **22:00** – Hostages are taken at the Bataclan.[43]

14 November:

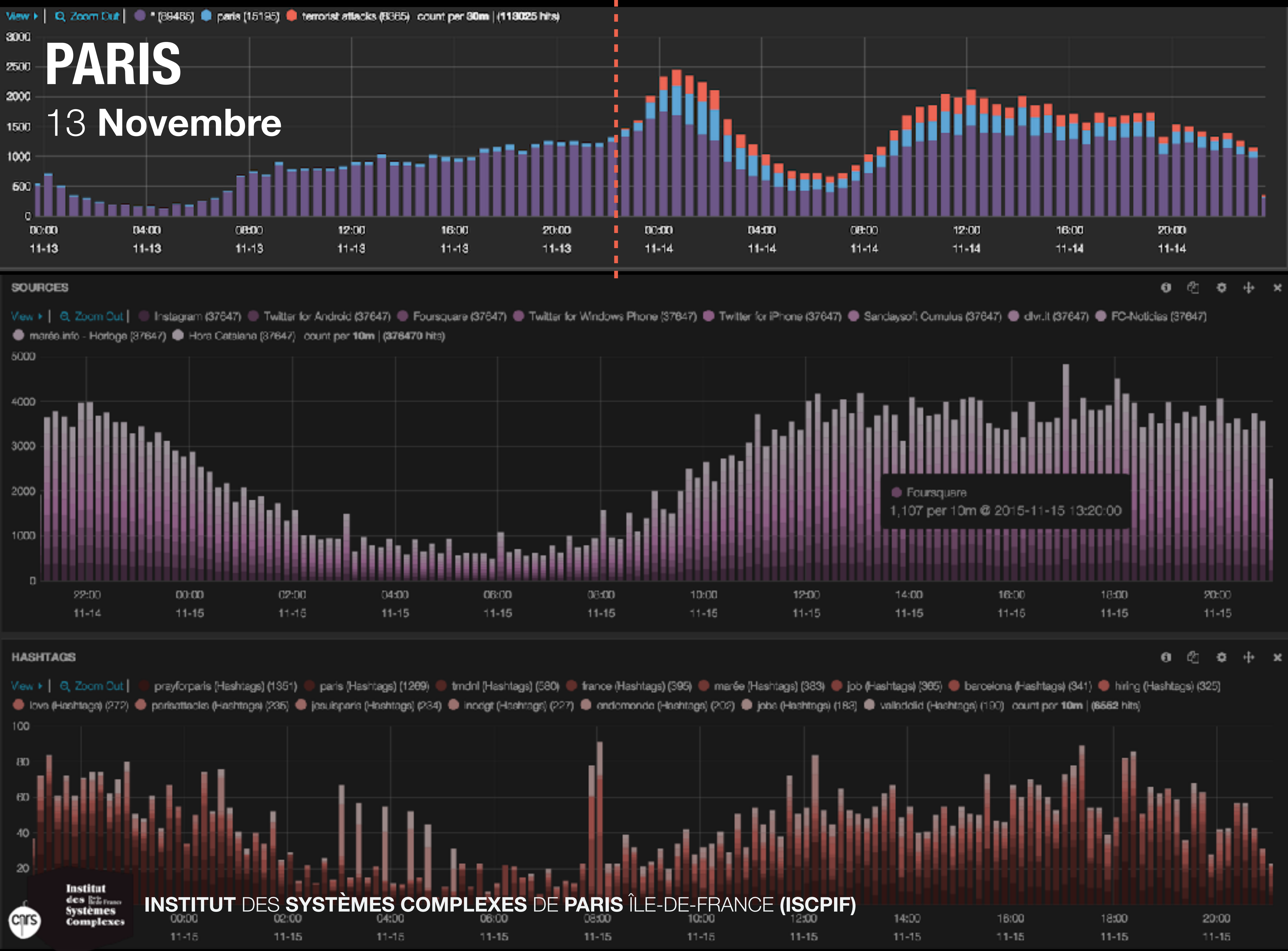
- **00:20** – Security forces enter the Bataclan.[7]
- **00:58** – French police end the siege on the Bataclan.[43]

All times are CET (UTC+1).

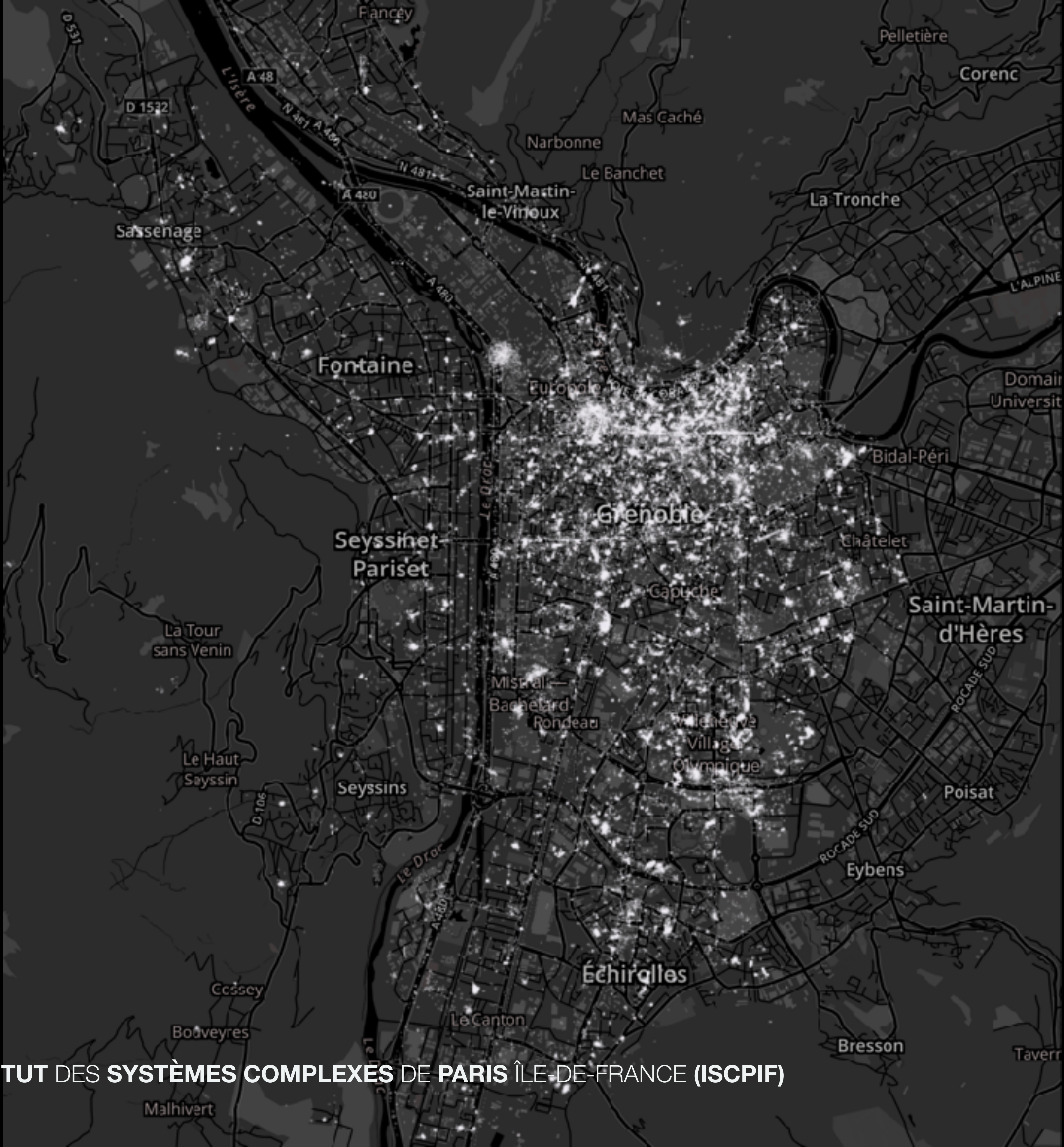


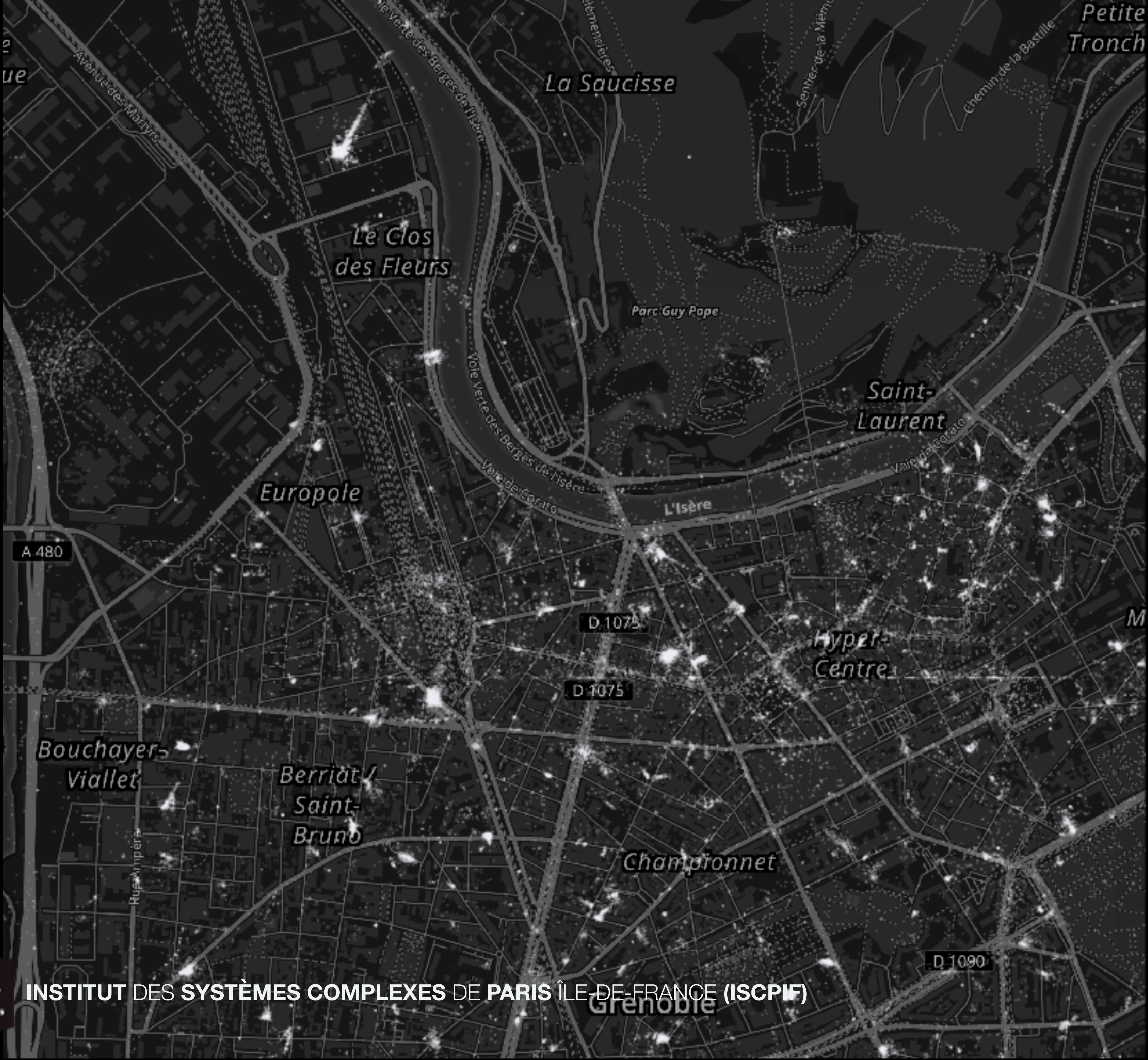
PARIS

13 Novembre


View ▸ | Zoom Out | ● * (89465) ● paris (15195) ● terrorist attacks (8365) count per 30m | (113025 hits)
3000
2500
2000
1500
1000
500
0
00:00 11-13
04:00 11-13
08:00 11-13
12:00 11-13
16:00 11-13
20:00 11-13
00:00 11-14
04:00 11-14
08:00 11-14
12:00 11-14
16:00 11-14
20:00 11-14
SOURCES
View ▸ | Zoom Out | ● Instagram (37647) ● Twitter for Android (37647) ● Foursquare (37647) ● Twitter for Windows Phone (37647) ● Twitter for iPhone (37647) ● Sandaysoft Cumulus (37647) ● dlvr.it (37647) ● FC-Noticias (37647) ● marée.info - Horloge (37647) ● Hora Catalana (37647) count per 10m | (376470 hits)
5000
4000
3000
2000
1000
0
Foursquare
1,107 per 10m @ 2015-11-15 13:20:00
22:00 11-14
00:00 11-15
02:00 11-15
04:00 11-15
06:00 11-15
08:00 11-15
10:00 11-15
12:00 11-15
14:00 11-15
16:00 11-15
18:00 11-15
20:00 11-15
HASHTAGS
View ▸ | Zoom Out | ● prayforparis (Hashtags) (1351) ● paris (Hashtags) (1269) ● tmdnl (Hashtags) (580) ● france (Hashtags) (395) ● marée (Hashtags) (383) ● job (Hashtags) (365) ● barcelona (Hashtags) (341) ● hiring (Hashtags) (325) ● love (Hashtags) (272) ● parisattacks (Hashtags) (235) ● jesuisparis (Hashtags) (234) ● inodgt (Hashtags) (227) ● endomondo (Hashtags) (202) ● jobs (Hashtags) (183) ● valladolid (Hashtags) (190) count per 10m | (6552 hits)
100
80
60
40
20
00:00 11-15
02:00 11-15
04:00 11-15
06:00 11-15
08:00 11-15
10:00 11-15
12:00 11-15
14:00 11-15
16:00 11-15
18:00 11-15
20:00 11-15

INSTITUT DES **SYSTÈMES COMPLEXES** DE **PARIS** ÎLE-DE-FRANCE **(ISCPIF)**


Fancey
D 531
A 48
L'Isère
D 1532
N 481
A 480
Pelletière
Corenc
Mas Caché
Narbonne
Le Banchet
Saint-Martin-le-Vinoux
La Tronche
Sassenage
L'ALPINE
Fontaine
Europole
Domaine Universit
Bidal-Péri
Le Drac
Grenoble
Seyssinet-Pariset
Châtelet
Capuche
Saint-Martin-d'Hères
La Tour sans Venin
Mistral — Bachelard
Rondeau
ROCADE SUD
Villeneuve Village Olympique
Le Haut Seyssin
Seyssins
Poisat
D 106
Eybens
Échirolles
Cossey
Le Canton
Bouveyres
Bresson
Taverr
Malhivert


La Saucisse
Petite Tronch
Le Clos des Fleurs
Parc Guy Pape
Saint-Laurent
Europole
L'Isère
A 480
D 1075
D 1075
Hyper-Centre
Bouchayer-Viallet
Berriat / Saint-Bruno
Championnet
D 1090
Grenoble
Avenue des Martyrs
Voie Verte des Berges de l'Isère
Voie de Corato
Sentier de la Mémoire
Chemin de la Bastille
Rue Ampère

MULTIVAC DATA SCIENCE LAB

MULTIVAC HADOOP CLUSTER

We have designed and implemented **Hadoop cluster** over more than 30 servers inside our private Cloud. This gives us Hadoop components such as **YARN**, **HDFS**, **Apache Spark**, **Apache Hive**, etc.

Multivac DSL offers a large-scale Hadoop cluster with over **900 vCores**, **1TB** of memory and more than **100TB** of distributed storage.

MULTIVAC HADOOP NOTEBOOKS

Multivac also offers interactive Hadoop notebooks by hosting multi-users/multi-tenants Apache Zeppelin and Hue.
Users can submit their codes and jobs over Multivac Hadoop Cluster by using Apache Spark interactive shell and Spark submit or Multivac hosted interactive notebooks in Scala, Java, Python, R and SQL.

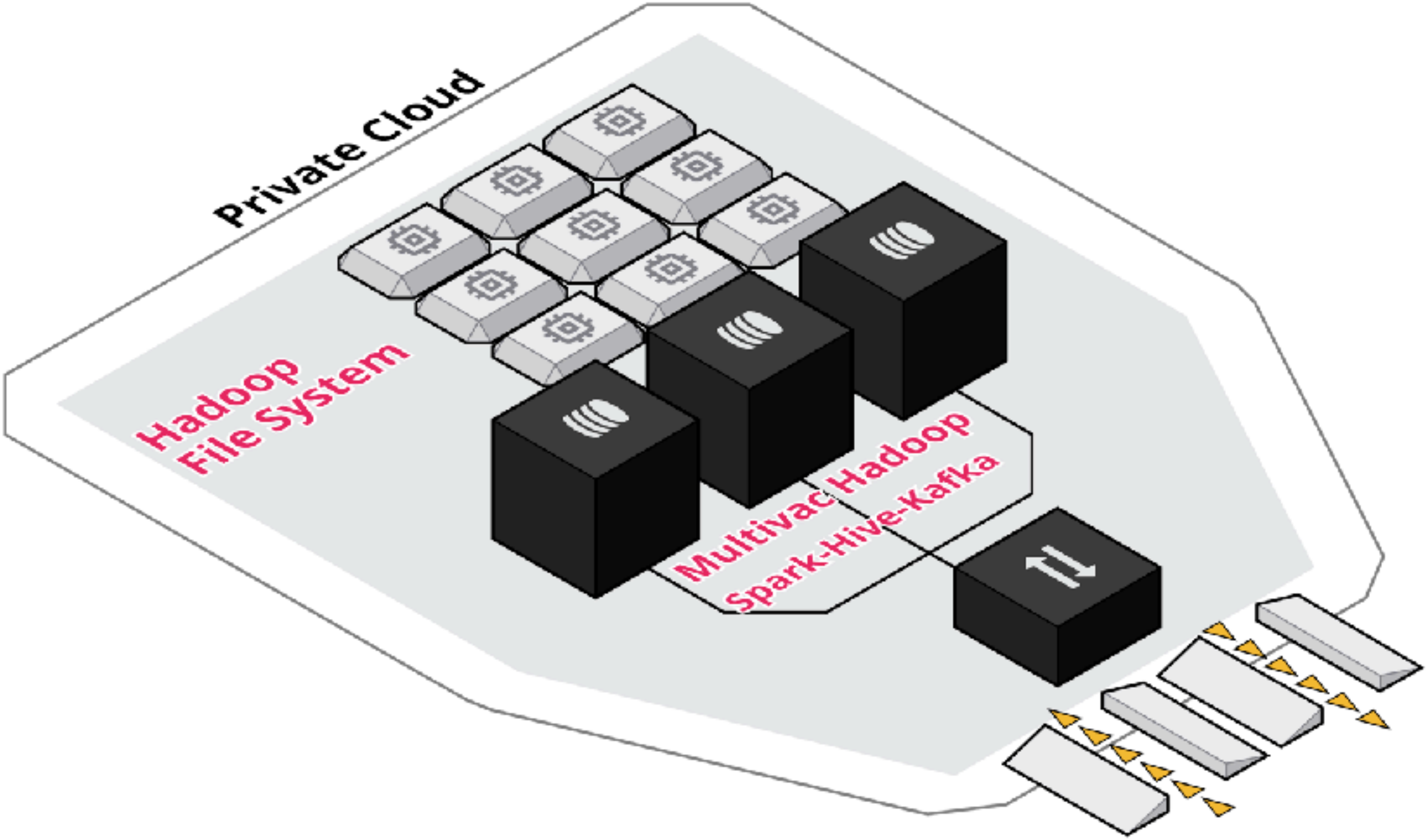
MULTIVAC HADOOP OPEN DATA

Multivac commits to make its important datasets to researchers and scientists over Multivac Hadoop Cluster in both format of RAW (JSON) and big SQL tables (Apache Hive). Users can run their jobs against Multivac Public Data easily!

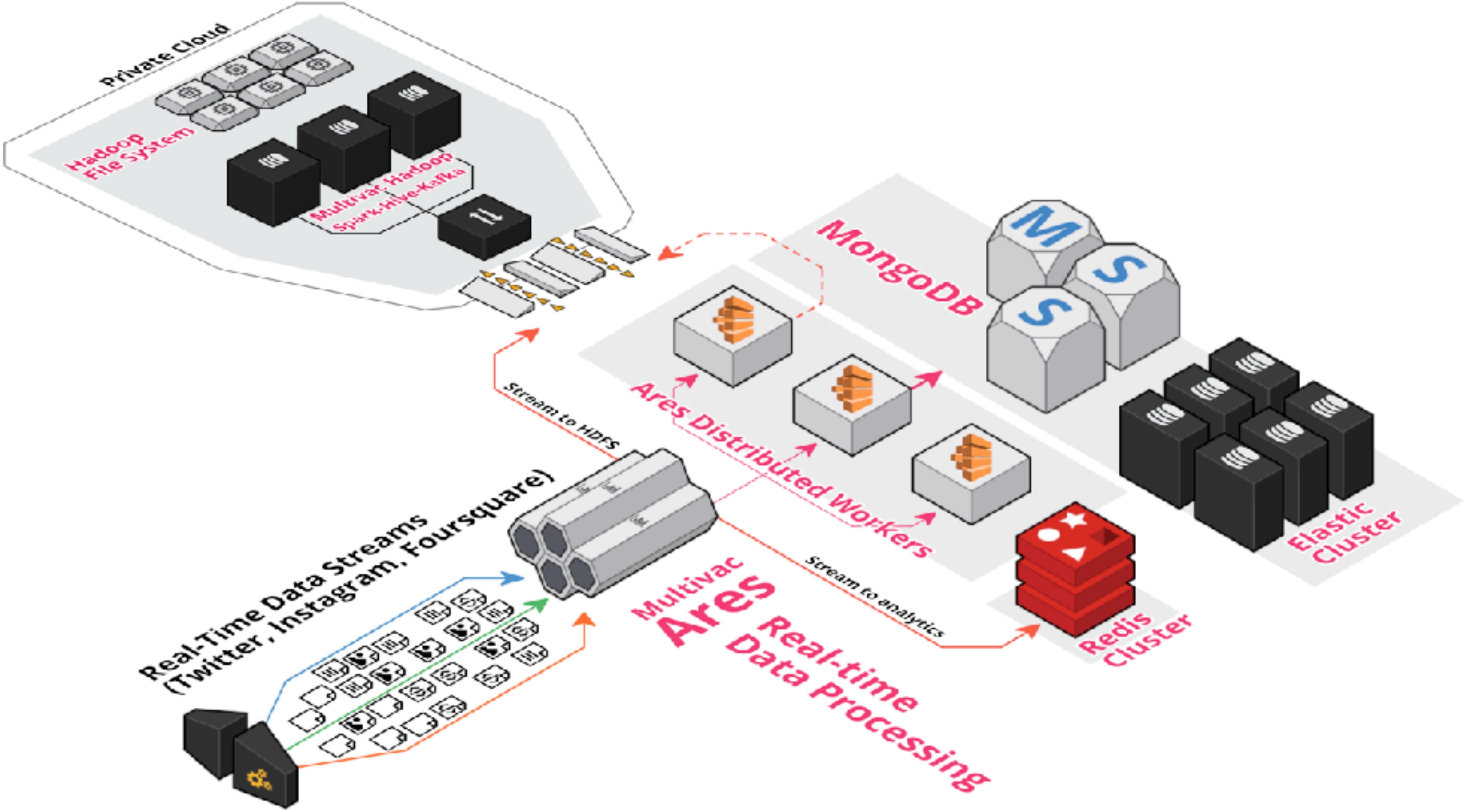
Multivac

Interactive Spark Notebooks

Big Data Infrastructure: Batch-processing & large-scale computations



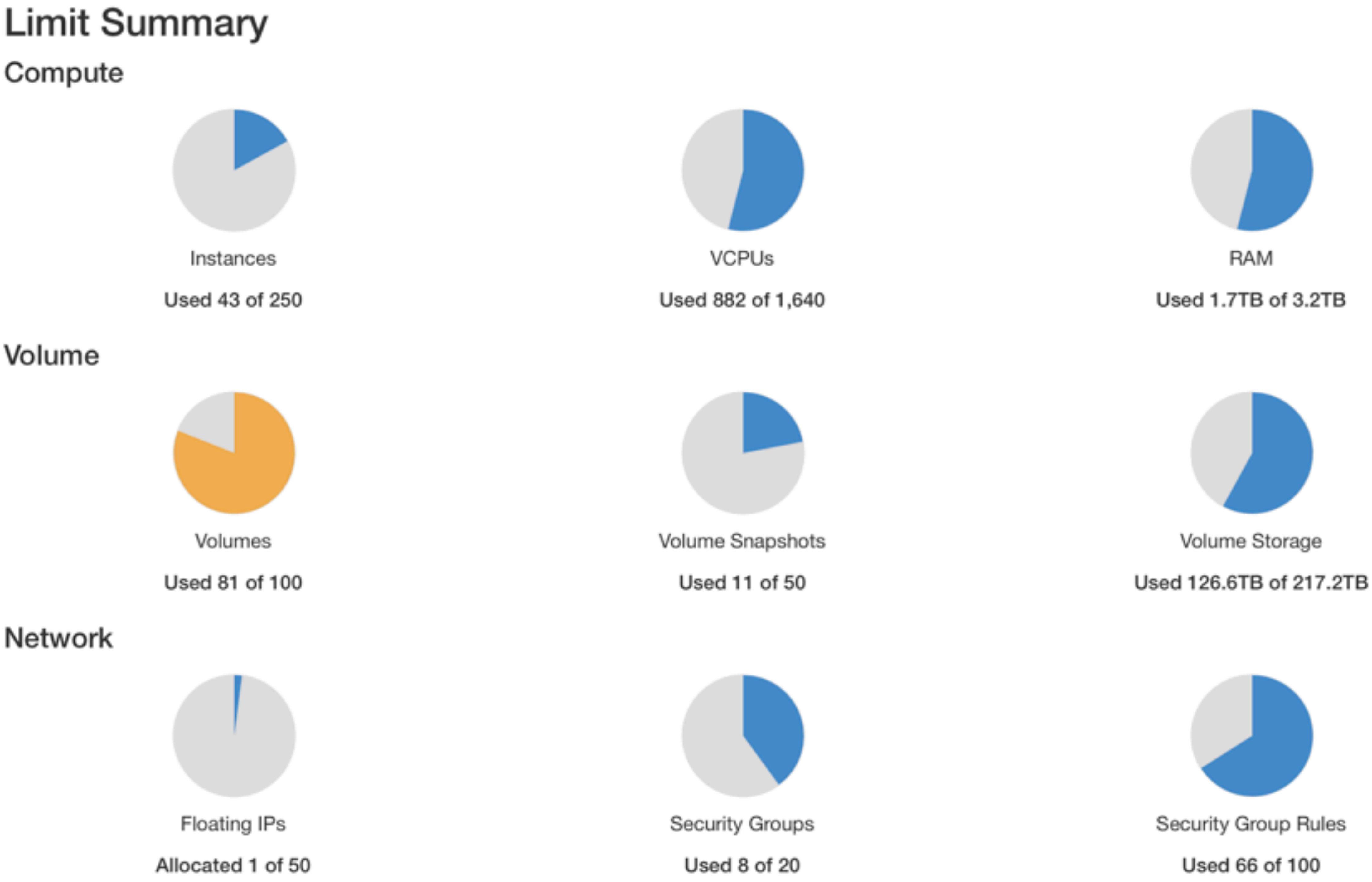
Big Data Infrastructure: Hybrid (Cloud & on-premise)


Private Cloud
Hadoop File System
Multivac Hadoop Spark-Hive-Kafka
MongoDB
Ares Distributed Workers
Elastic Cluster
Stream to HDFS
Stream to analytics
Real-Time Data Streams (Twitter, Instagram, Foursquare)
Multivac Ares Real-time Data Processing
Redis Cluster

INSTITUT DES **SYSTÈMES COMPLEXES** DE **PARIS** ÎLE-DE-FRANCE **(ISCPIF)**

https://multivacplatform.org/

Big Data Infrastructure: Batch-processing & large-scale computations



Limit Summary

Compute

Instances

Used 43 of 250

VCPUs

Used 882 of 1,640

RAM

Used 1.7TB of 3.2TB

Volume

Volumes

Used 81 of 100

Volume Snapshots

Used 11 of 50

Volume Storage

Used 126.6TB of 217.2TB

Network

Floating IPs

Allocated 1 of 50

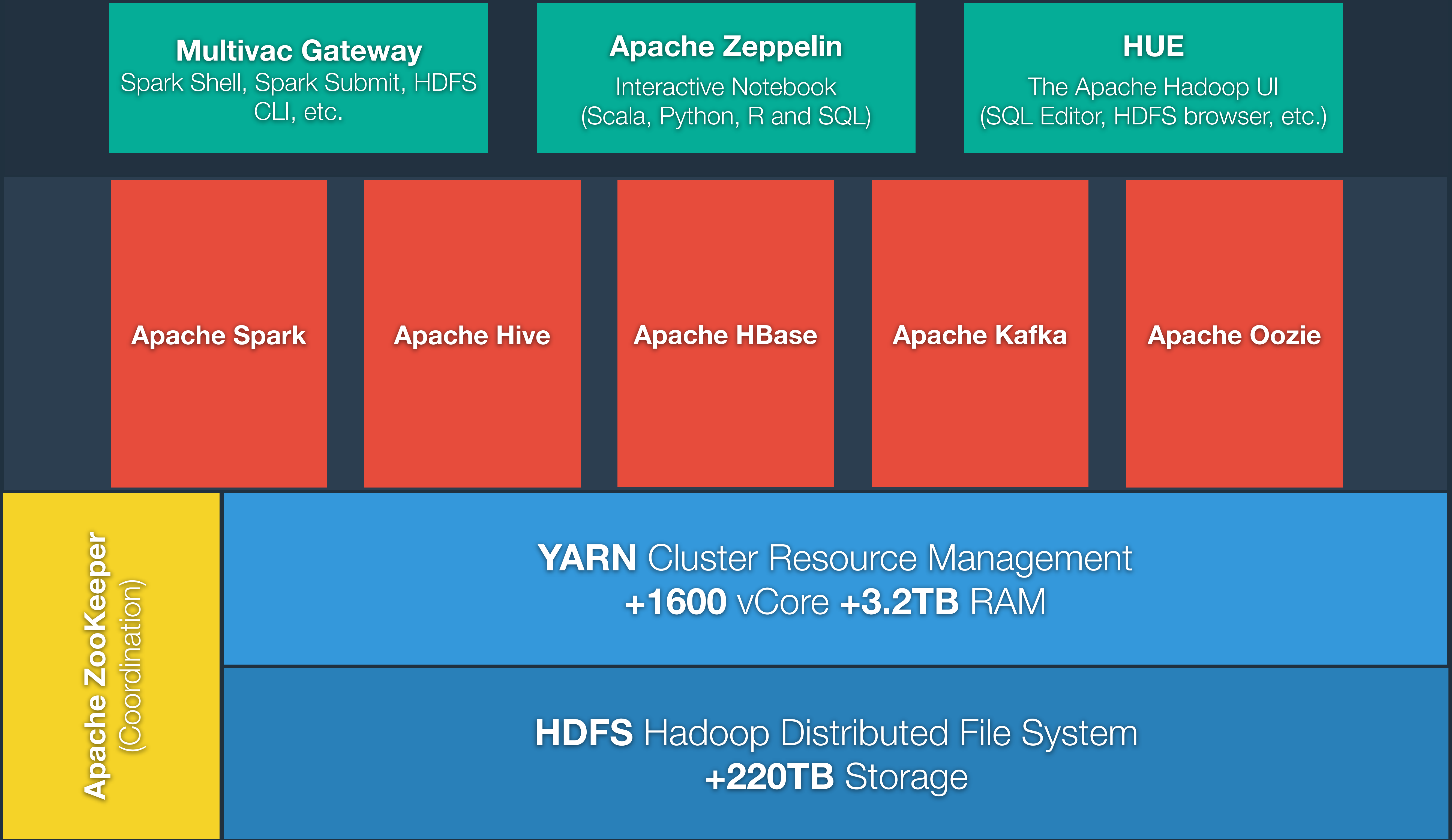
Security Groups

Used 8 of 20

Security Group Rules

Used 66 of 100

Multivac Data Science Lab



Multivac Gateway
Spark Shell, Spark Submit, HDFS CLI, etc.

Apache Zeppelin
Interactive Notebook
(Scala, Python, R and SQL)

HUE
The Apache Hadoop UI
(SQL Editor, HDFS browser, etc.)

Apache Spark

Apache Hive

Apache HBase

Apache Kafka

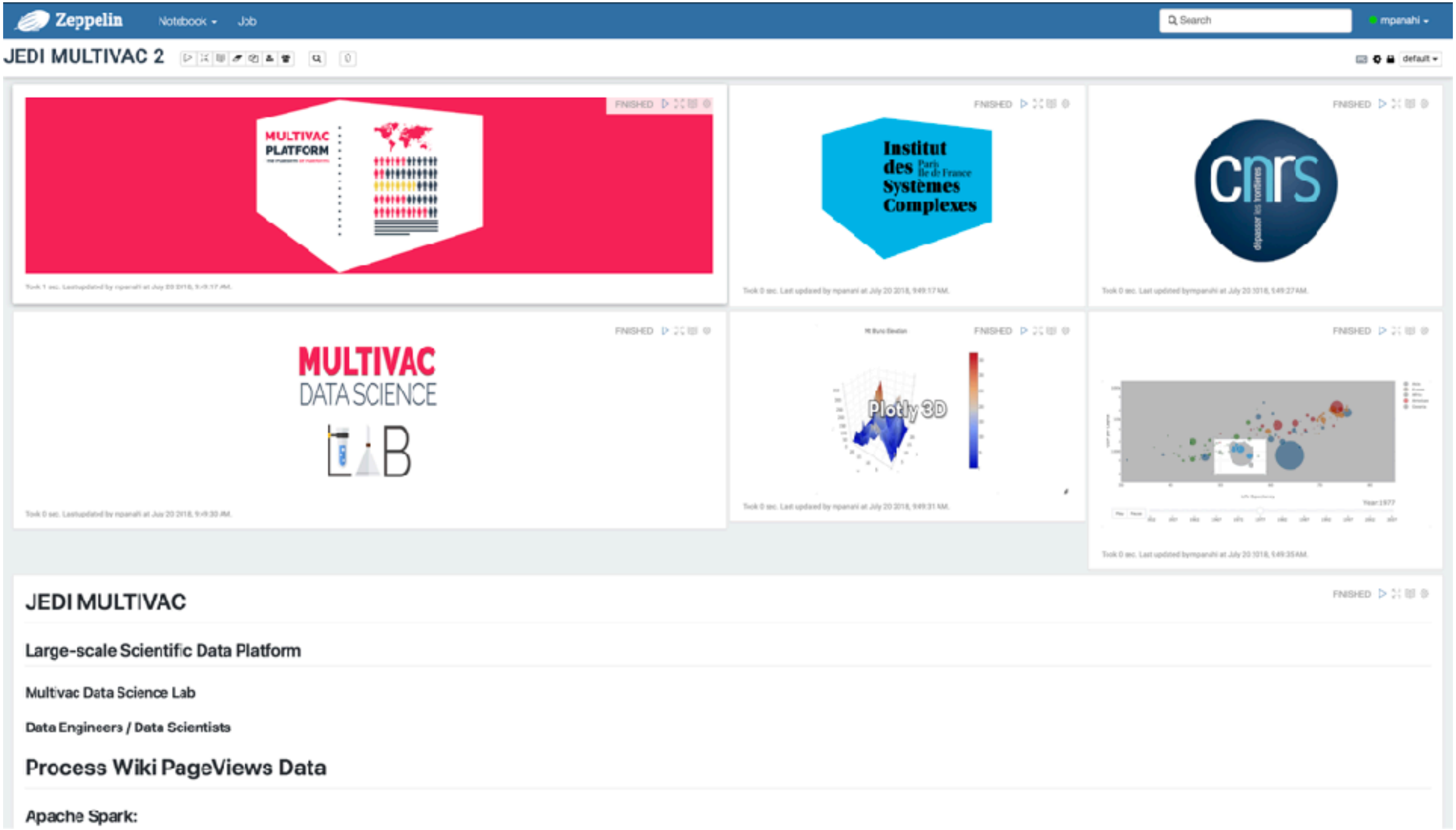
Apache Oozie

Apache ZooKeeper
(Coordination)

YARN Cluster Resource Management
+1600 vCore **+3.2TB** RAM

HDFS Hadoop Distributed File System
+220TB Storage

Apache **Zeppelin**



JEDI MULTIVAC

Large-scale Scientific Data Platform

Multivac Data Science Lab

Data Engineers / Data Scientists

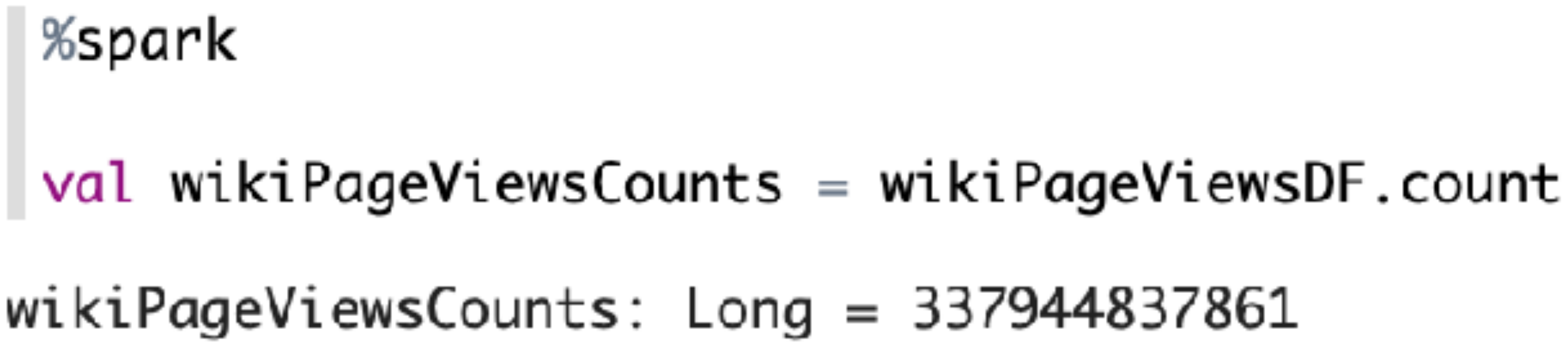
Process Wiki PageViews Data

Apache Spark:

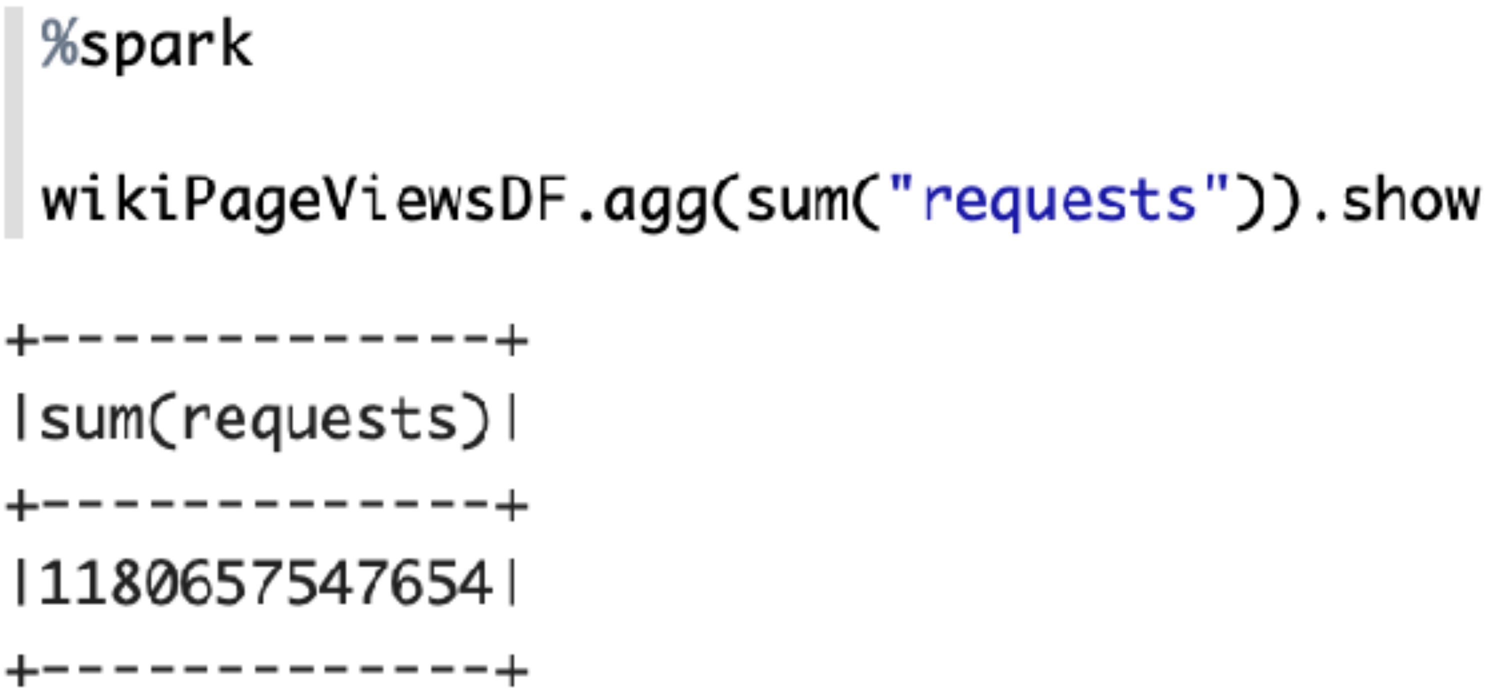
Wikimedia Project

Wikipedia hourly-generated page views!

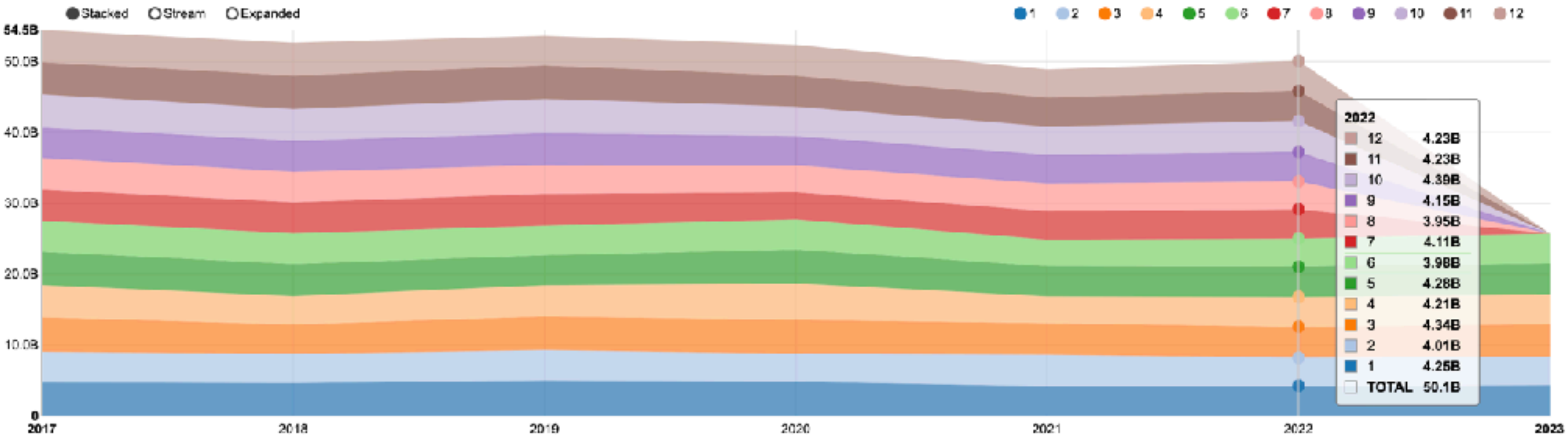
https://dumps.wikimedia.org/other/pageviews/



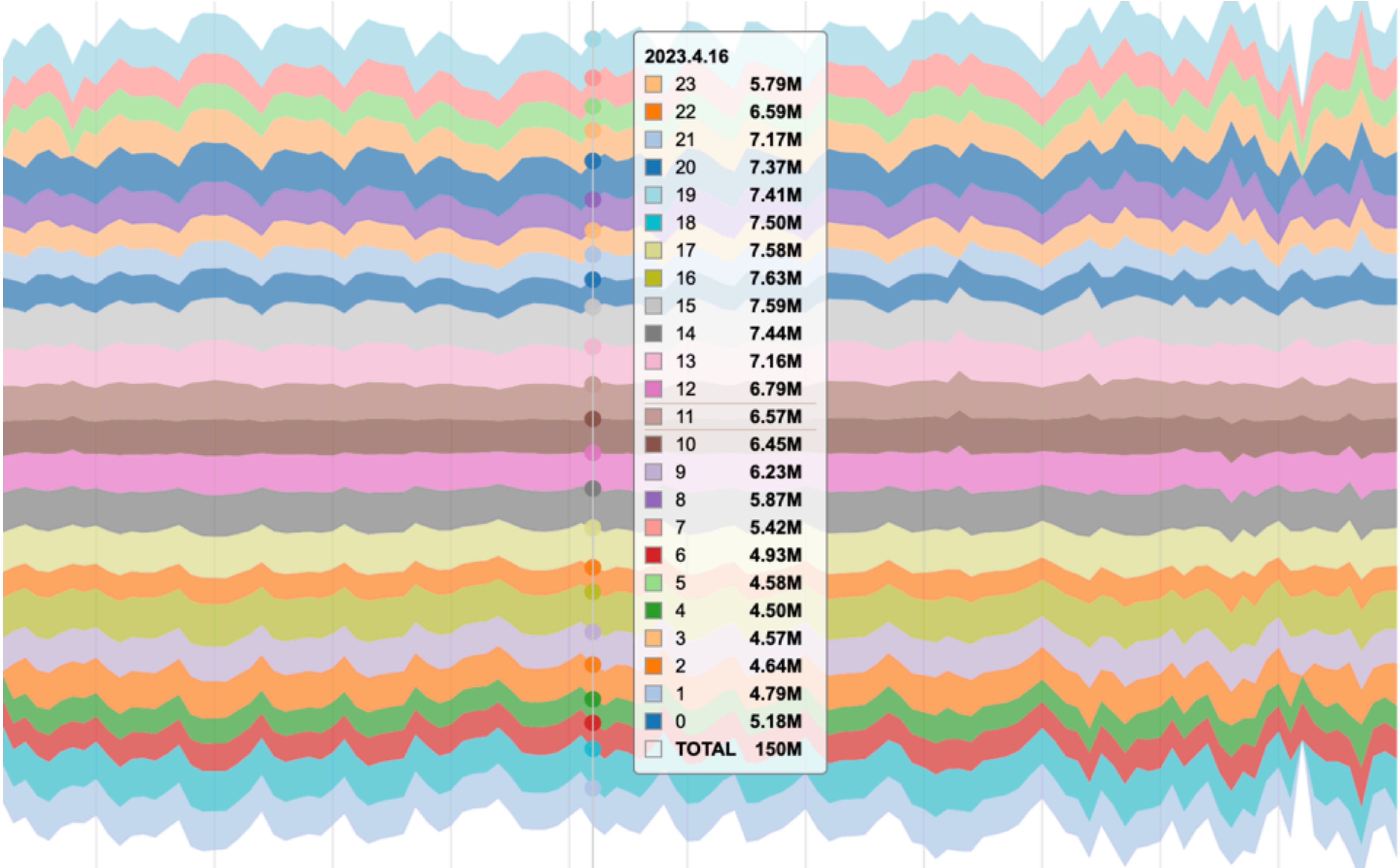
```
%spark

val wikiPageViewsCounts = wikiPageViewsDF.count

wikiPageViewsCounts: Long = 337944837861
```



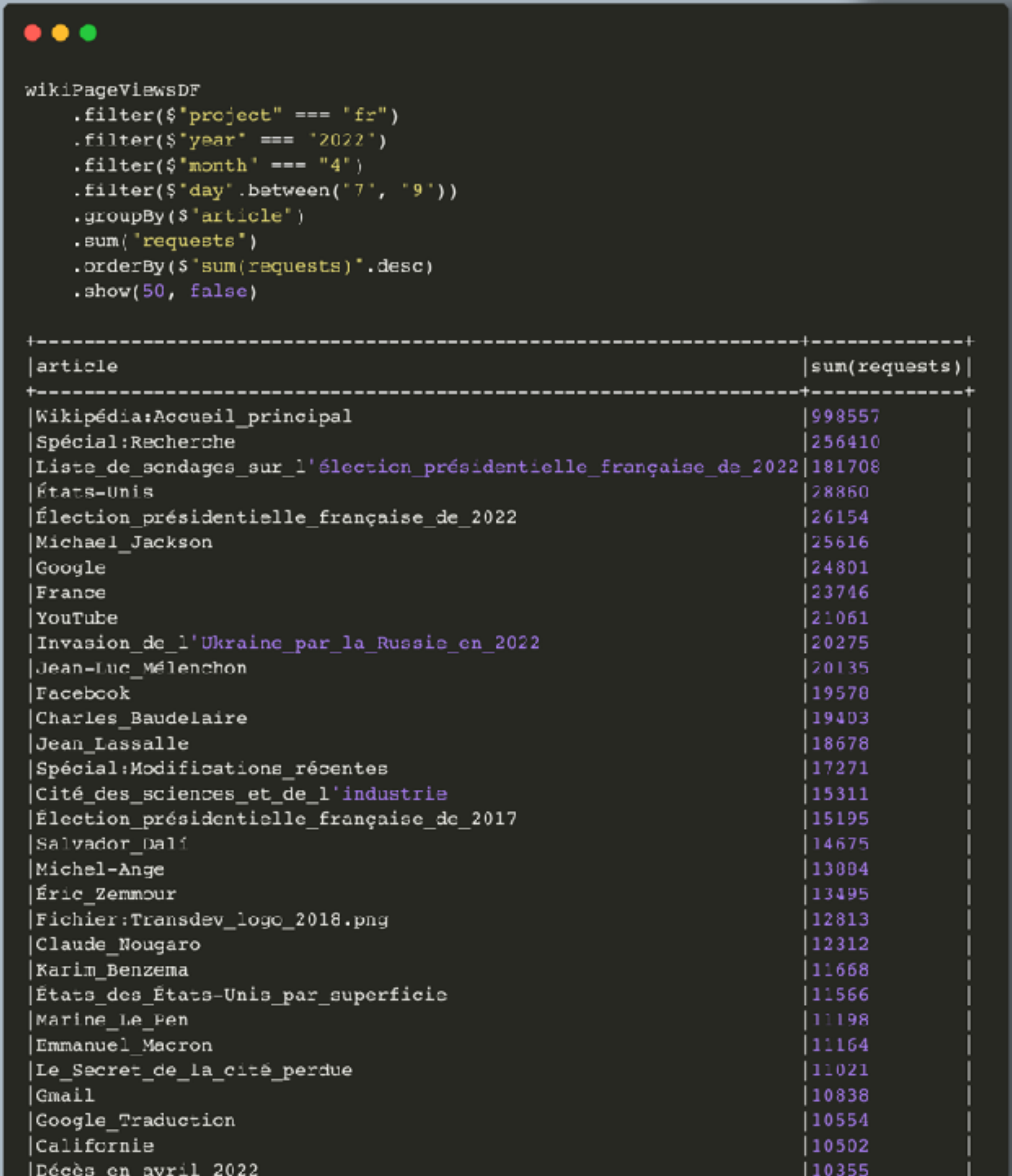
```
%spark

wikiPageViewsDF.agg(sum("requests")).show

+-------------+
|sum(requests)|
+-------------+
|1180657547654|
+-------------+
```



2022	
12	4.23B
11	4.23B
10	4.39B
9	4.15B
8	3.95B
7	4.11B
6	3.98B
5	4.28B
4	4.21B
3	4.34B
2	4.01B
1	4.25B
TOTAL	50.1B


2023.4.16
23 5.79M
22 6.59M
21 7.17M
20 7.37M
19 7.41M
18 7.50M
17 7.58M
16 7.63M
15 7.59M
14 7.44M
13 7.16M
12 6.79M
11 6.57M
10 6.45M
9 6.23M
8 5.87M
7 5.42M
6 4.93M
5 4.58M
4 4.50M
3 4.57M
2 4.64M
1 4.79M
0 5.18M
TOTAL 150M

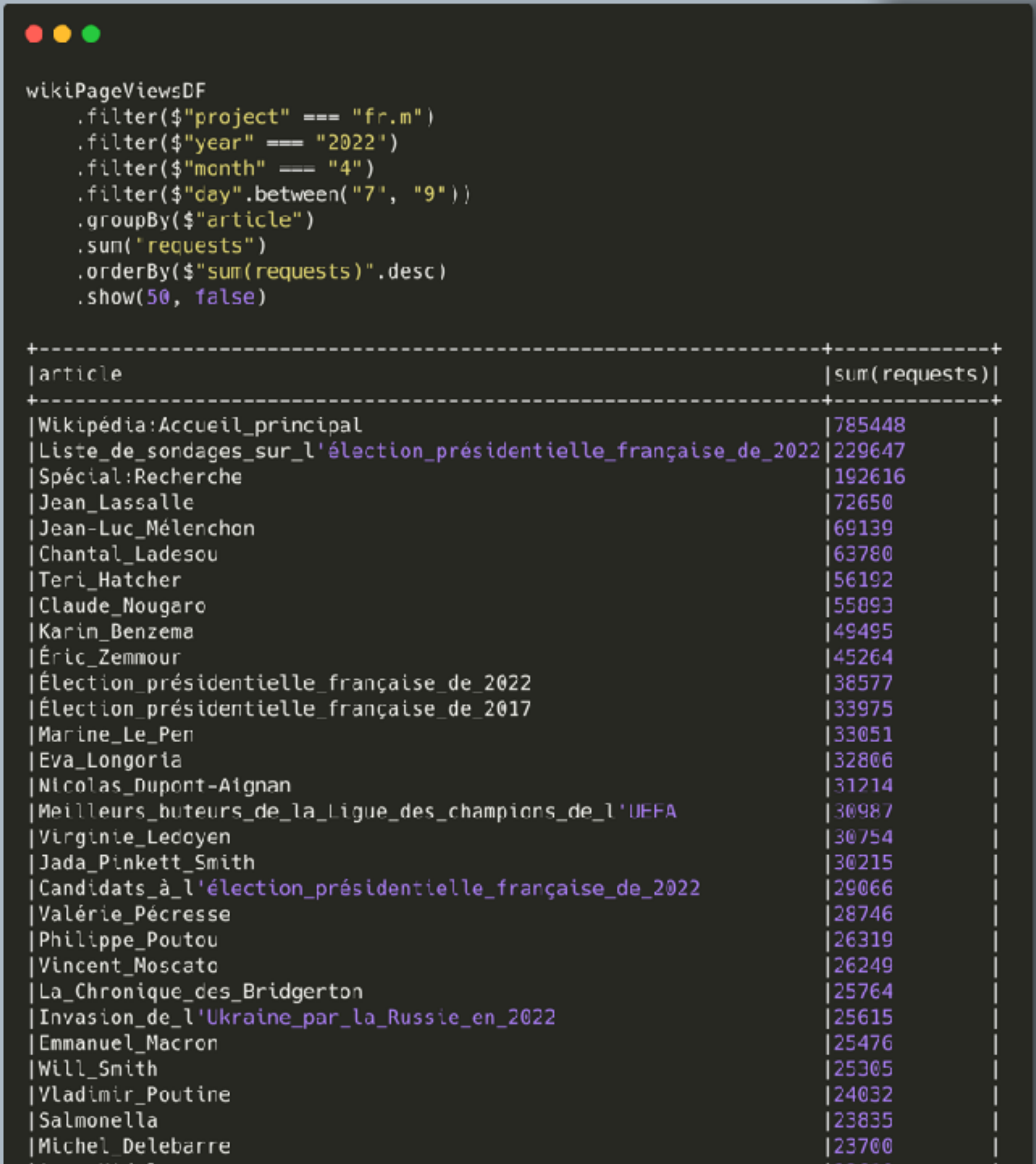
Pulse of **Internet**



```
wikiPageViewsDF
    .filter($"project" === "fr")
    .filter($"year" === "2022")
    .filter($"month" === "4")
    .filter($"day".between("7", "9"))
    .groupBy($"article")
    .sum("requests")
    .orderBy($"sum(requests)".desc)
    .show(50, false)

+-----------------------------------------------------------------+-------------+
|article                                                          |sum(requests)|
+-----------------------------------------------------------------+-------------+
|Wikipédia:Accueil_principal                                      |998557       |
|Spécial:Recherche                                                |256410       |
|Liste_de_sondages_sur_l'élection_présidentielle_française_de_2022|181708       |
|États-Unis                                                       |28860        |
|Élection_présidentielle_française_de_2022                        |26154        |
|Michael_Jackson                                                  |25616        |
|Google                                                           |24801        |
|France                                                           |23746        |
|YouTube                                                          |21061        |
|Invasion_de_l'Ukraine_par_la_Russie_en_2022                      |20275        |
|Jean-Luc_Mélenchon                                               |20135        |
|Facebook                                                         |19578        |
|Charles_Baudelaire                                               |19403        |
|Jean_Lassalle                                                    |18678        |
|Spécial:Modifications_récentes                                   |17271        |
|Cité_des_sciences_et_de_l'industrie                              |15311        |
|Élection_présidentielle_française_de_2017                        |15195        |
|Salvador_Dalí                                                    |14675        |
|Michel-Ange                                                      |13884        |
|Éric_Zemmour                                                     |13495        |
|Fichier:Transdev_logo_2018.png                                   |12813        |
|Claude_Nougaro                                                   |12312        |
|Karim_Benzema                                                    |11668        |
|États_des_États-Unis_par_superficie                              |11566        |
|Marine_Le_Pen                                                    |11198        |
|Emmanuel_Macron                                                  |11164        |
|Le_Secret_de_la_cité_perdue                                      |11021        |
|Gmail                                                            |10838        |
|Google_Traduction                                                |10554        |
|Californie                                                       |10502        |
|Décès_en_avril_2022                                              |10355        |
```

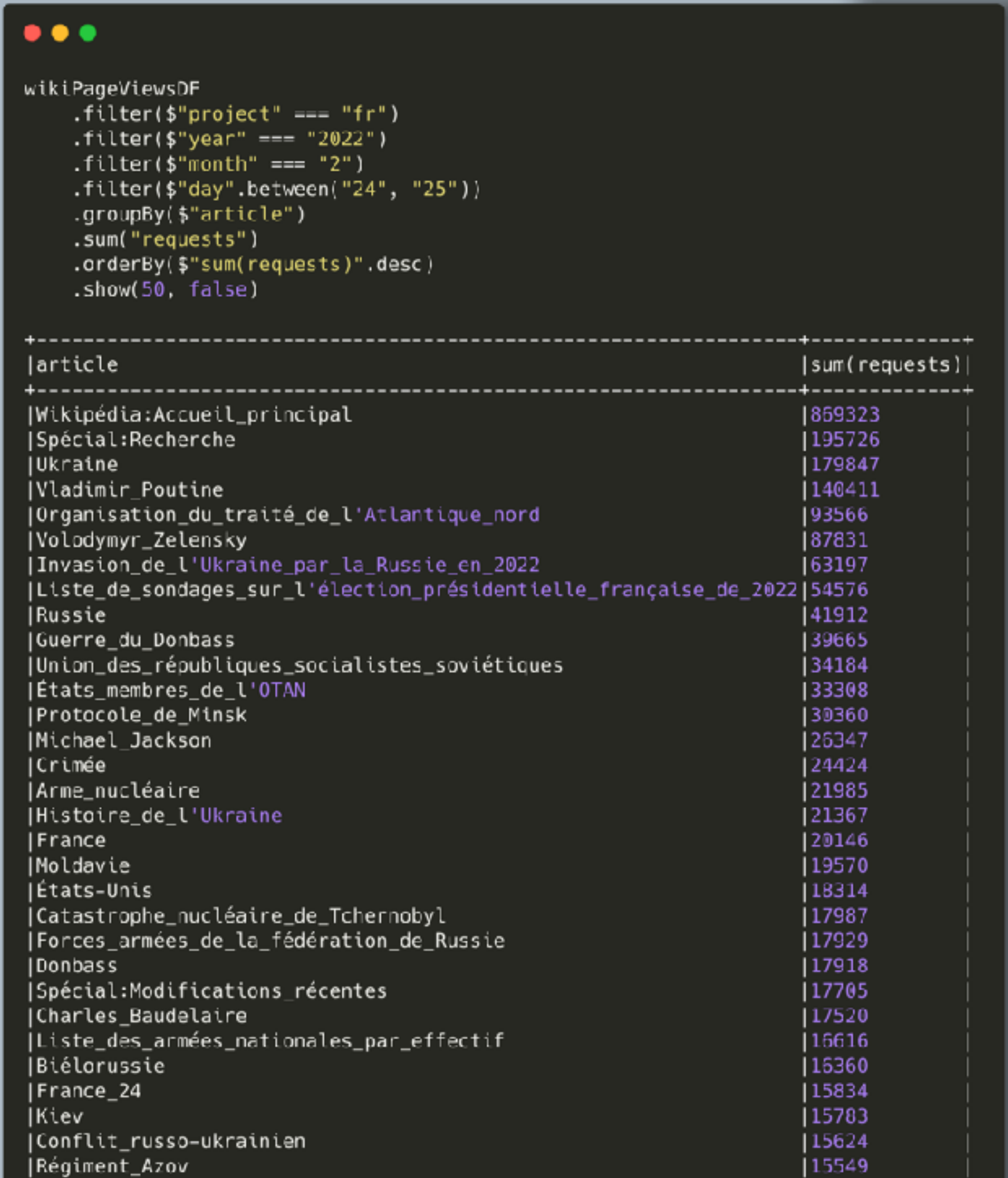
Institut des Systèmes Complexes

Pulse of **Internet**



```
wikiPageViewsDF
    .filter($"project" === "fr.m")
    .filter($"year" === "2022")
    .filter($"month" === "4")
    .filter($"day".between("7", "9"))
    .groupBy($"article")
    .sum("requests")
    .orderBy($"sum(requests)".desc)
    .show(50, false)

+----------------------------------------------------------------+--------------+
|article                                                         |sum(requests)|
+----------------------------------------------------------------+--------------+
|Wikipédia:Accueil_principal                                     |785448        |
|Liste_de_sondages_sur_l'élection_présidentielle_française_de_2022|229647        |
|Spécial:Recherche                                               |192616        |
|Jean_Lassalle                                                   |72650         |
|Jean-Luc_Mélenchon                                              |69139         |
|Chantal_Ladesou                                                 |63780         |
|Teri_Hatcher                                                    |56192         |
|Claude_Nougaro                                                  |55893         |
|Karim_Benzema                                                   |49495         |
|Éric_Zemmour                                                    |45264         |
|Élection_présidentielle_française_de_2022                       |38577         |
|Élection_présidentielle_française_de_2017                       |33975         |
|Marine_Le_Pen                                                   |33051         |
|Eva_Longoria                                                    |32806         |
|Nicolas_Dupont-Aignan                                           |31214         |
|Meilleurs_buteurs_de_la_Ligue_des_champions_de_l'UEFA           |30987         |
|Virginie_Ledoyen                                                |30754         |
|Jada_Pinkett_Smith                                              |30215         |
|Candidats_à_l'élection_présidentielle_française_de_2022         |29066         |
|Valérie_Pécresse                                                |28746         |
|Philippe_Poutou                                                 |26319         |
|Vincent_Moscato                                                 |26249         |
|La_Chronique_des_Bridgerton                                     |25764         |
|Invasion_de_l'Ukraine_par_la_Russie_en_2022                     |25615         |
|Emmanuel_Macron                                                 |25476         |
|Will_Smith                                                      |25305         |
|Vladimir_Poutine                                                |24032         |
|Salmonella                                                      |23835         |
|Michel_Delebarre                                                |23700         |
```

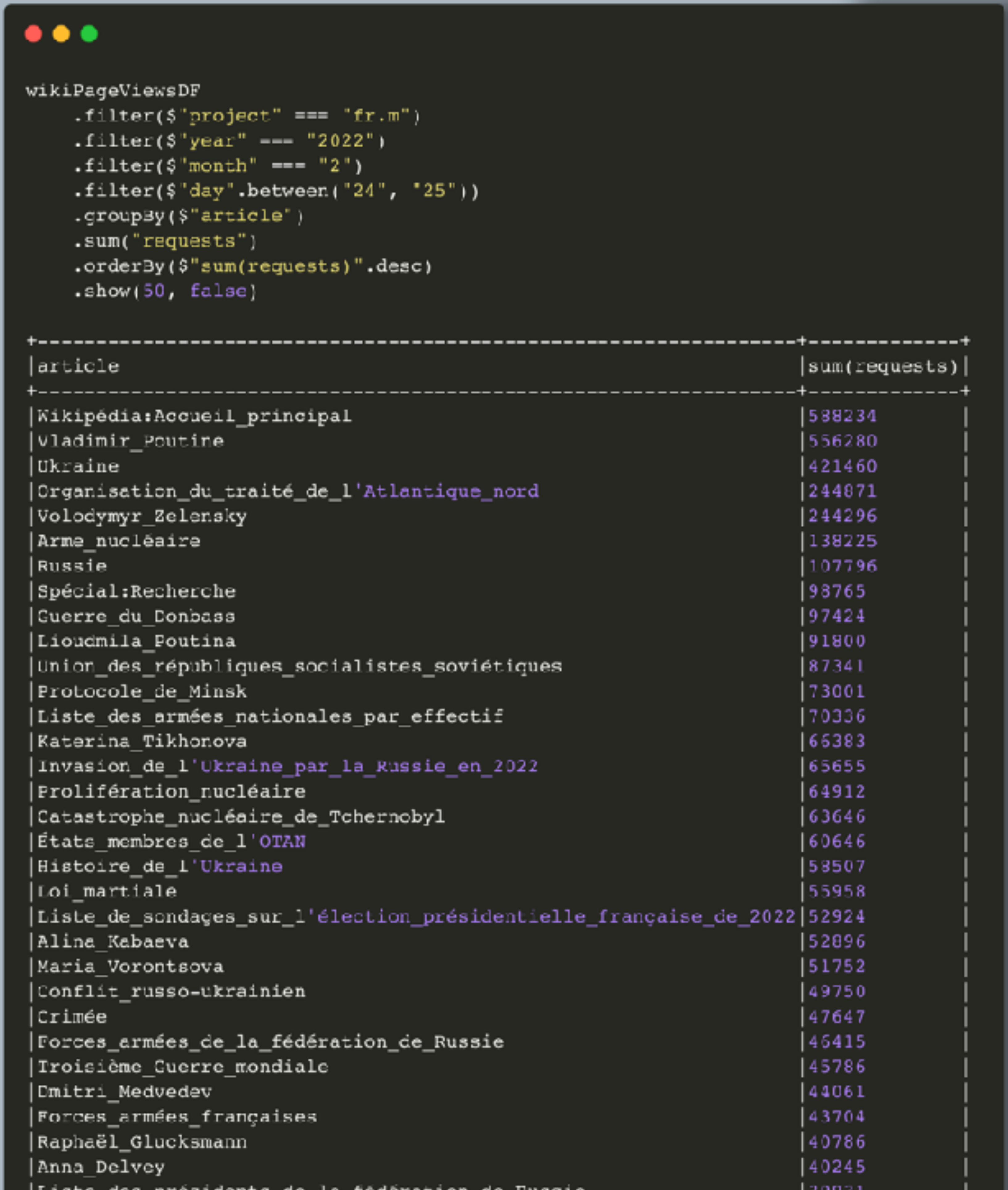
Pulse of **Internet**



```
wikiPageViewsDF
    .filter($"project" === "fr")
    .filter($"year" === "2022")
    .filter($"month" === "2")
    .filter($"day".between("24", "25"))
    .groupBy($"article")
    .sum("requests")
    .orderBy($"sum(requests)".desc)
    .show(50, false)

+-----------------------------------------------------------------+-------------+
|article                                                          |sum(requests)|
+-----------------------------------------------------------------+-------------+
|Wikipédia:Accueil_principal                                      |869323       |
|Spécial:Recherche                                                |195726       |
|Ukraine                                                          |179847       |
|Vladimir_Poutine                                                 |140411       |
|Organisation_du_traité_de_l'Atlantique_nord                      |93566        |
|Volodymyr_Zelensky                                               |87831        |
|Invasion_de_l'Ukraine_par_la_Russie_en_2022                      |63197        |
|Liste_de_sondages_sur_l'élection_présidentielle_française_de_2022|54576        |
|Russie                                                           |41912        |
|Guerre_du_Donbass                                                |39665        |
|Union_des_républiques_socialistes_soviétiques                    |34184        |
|États_membres_de_l'OTAN                                          |33308        |
|Protocole_de_Minsk                                               |30360        |
|Michael_Jackson                                                  |26347        |
|Crimée                                                           |24424        |
|Arme_nucléaire                                                   |21985        |
|Histoire_de_l'Ukraine                                            |21367        |
|France                                                           |20146        |
|Moldavie                                                         |19570        |
|États-Unis                                                       |18314        |
|Catastrophe_nucléaire_de_Tchernobyl                              |17987        |
|Forces_armées_de_la_fédération_de_Russie                         |17929        |
|Donbass                                                          |17918        |
|Spécial:Modifications_récentes                                   |17705        |
|Charles_Baudelaire                                               |17520        |
|Liste_des_armées_nationales_par_effectif                         |16616        |
|Biélorussie                                                      |16360        |
|France_24                                                        |15834        |
|Kiev                                                             |15783        |
|Conflit_russo-ukrainien                                          |15624        |
|Régiment_Azov                                                    |15549        |
```

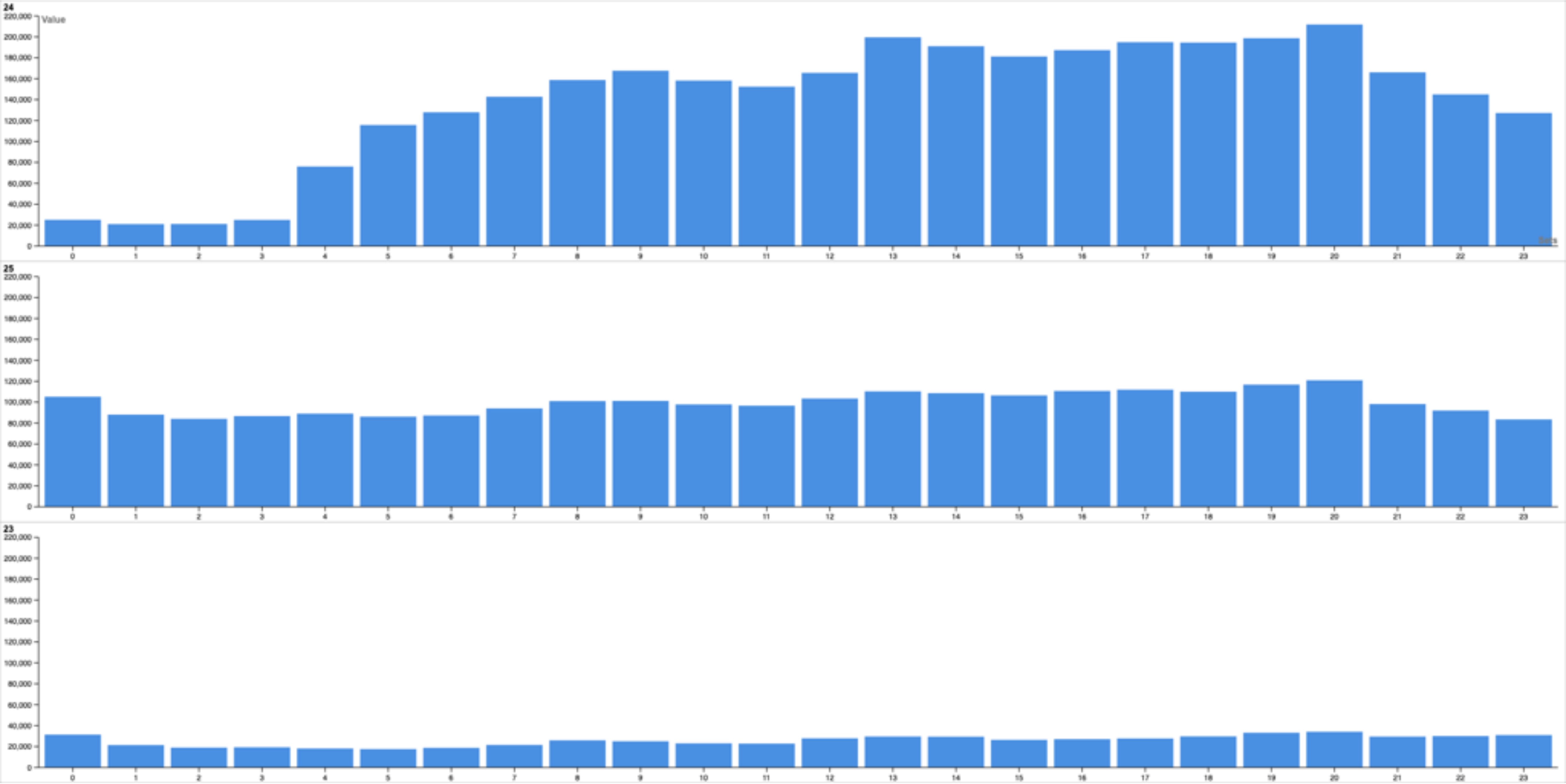
Institut des Systèmes Complexes

Pulse of **Internet**

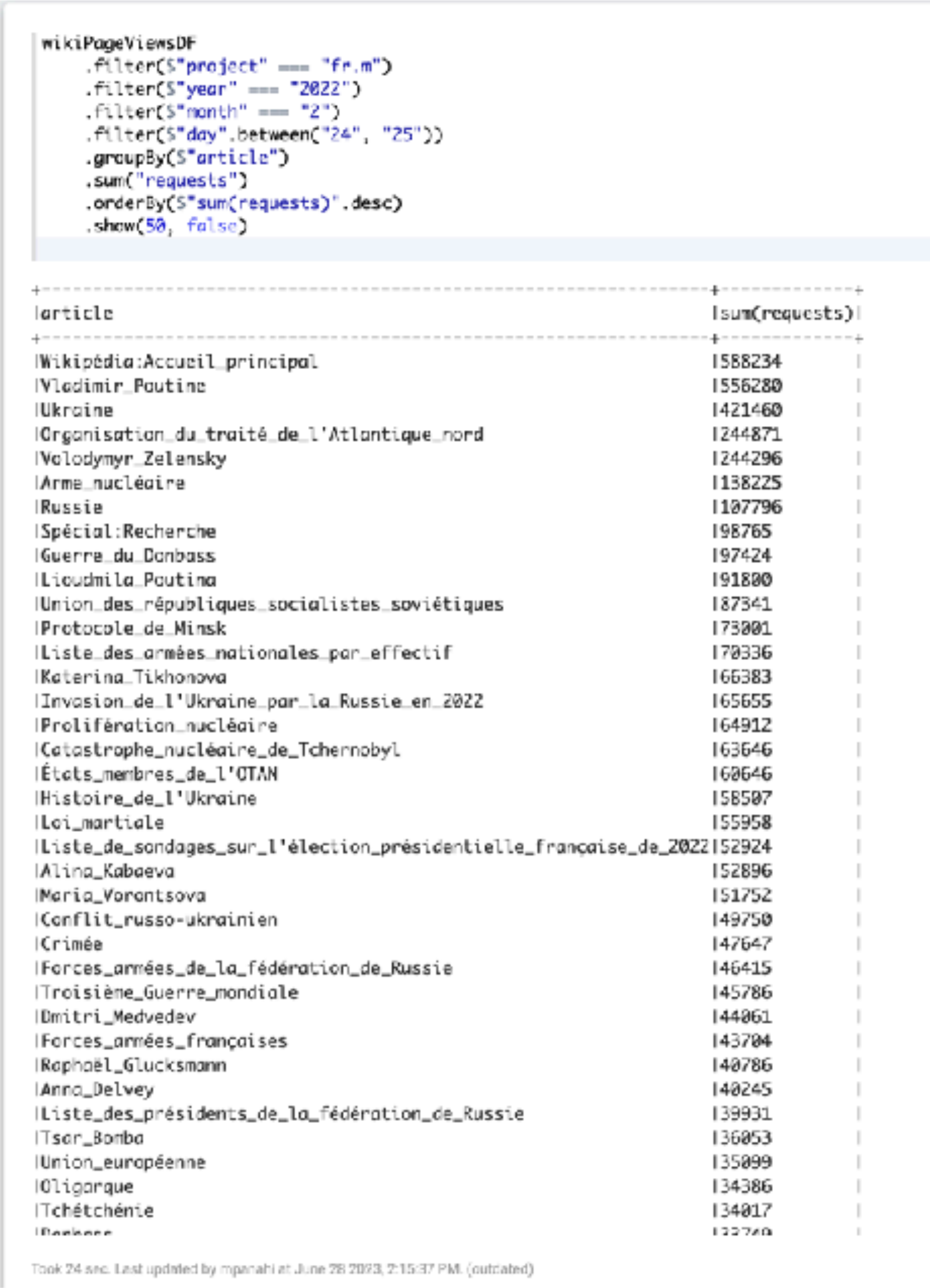


```
wikiPageViewsDF
    .filter($"project" === "fr.m")
    .filter($"year" === "2022")
    .filter($"month" === "2")
    .filter($"day".between("24", "25"))
    .groupBy($"article")
    .sum("requests")
    .orderBy($"sum(requests)".desc)
    .show(50, false)

+-----------------------------------------------------------------+------------+
|article                                                          |sum(requests)|
+-----------------------------------------------------------------+------------+
|Wikipédia:Accueil_principal                                      |588234      |
|Vladimir_Poutine                                                 |556280      |
|Ukraine                                                          |421460      |
|Organisation_du_traité_de_l'Atlantique_nord                      |244871      |
|Volodymyr_Zelensky                                               |244296      |
|Arme_nucléaire                                                   |138225      |
|Russie                                                           |107796      |
|Spécial:Recherche                                                |98765       |
|Guerre_du_Donbass                                                |97424       |
|Lioudmila_Poutina                                                |91800       |
|Union_des_républiques_socialistes_soviétiques                    |87341       |
|Protocole_de_Minsk                                               |73001       |
|Liste_des_armées_nationales_par_effectif                         |70336       |
|Katerina_Tikhonova                                               |66383       |
|Invasion_de_l'Ukraine_par_la_Russie_en_2022                      |65655       |
|Prolifération_nucléaire                                          |64912       |
|Catastrophe_nucléaire_de_Tchernobyl                              |63646       |
|États_membres_de_l'OTAN                                          |60646       |
|Histoire_de_l'Ukraine                                            |58507       |
|Loi_martiale                                                     |55958       |
|Liste_de_sondages_sur_l'élection_présidentielle_française_de_2022|52924       |
|Alina_Kabaeva                                                    |52896       |
|Maria_Vorontsova                                                 |51752       |
|Conflit_russo-ukrainien                                          |49750       |
|Crimée                                                           |47647       |
|Forces_armées_de_la_fédération_de_Russie                         |46415       |
|Troisième_Guerre_mondiale                                        |45786       |
|Dmitri_Medvedev                                                  |44061       |
|Forces_armées_françaises                                         |43704       |
|Raphaël_Glucksmann                                               |40786       |
|Anna_Delvey                                                      |40245       |
|Liste_des_présidents_de_la_fédération_de_Russie
```

Ukraine - 23rd, 24th, and 25th February



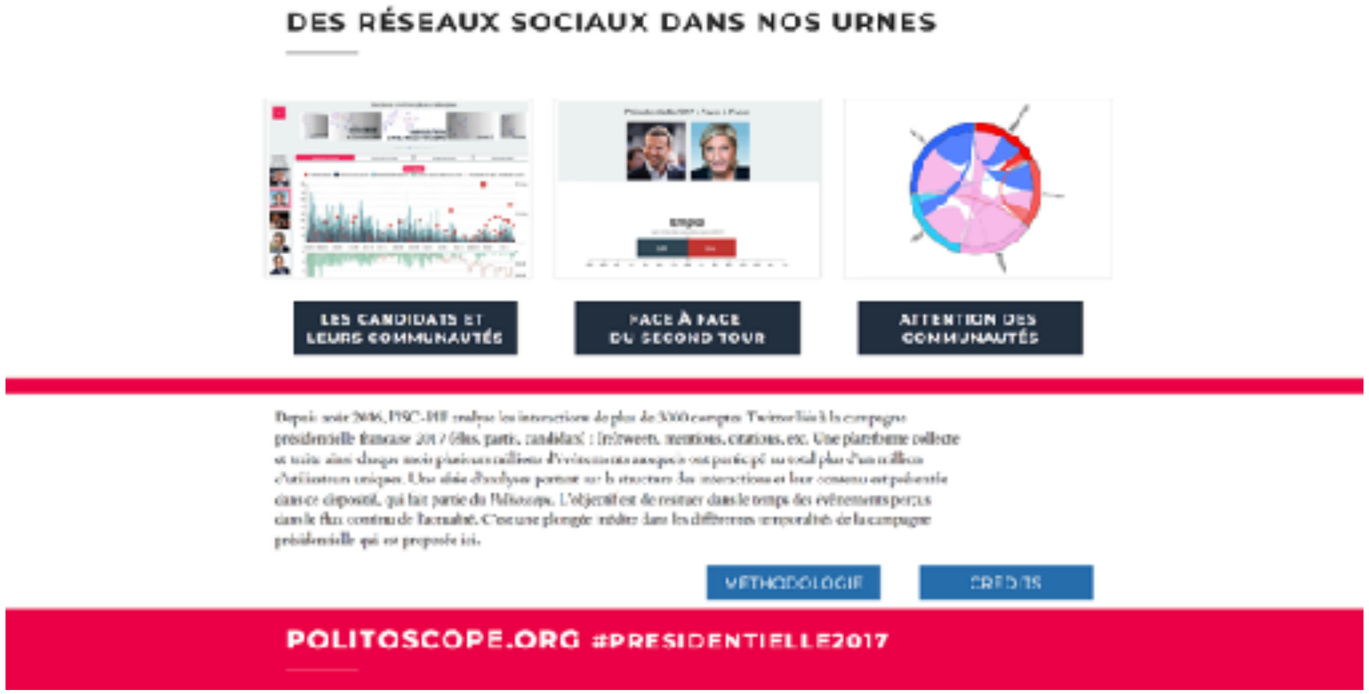
337 billion records - less **than minute**



```
wikiPageViewsDF
    .filter($"project" === "fr.m")
    .filter($"year" === "2022")
    .filter($"month" === "2")
    .filter($"day".between("24", "25"))
    .groupBy($"article")
    .sum("requests")
    .orderBy($"sum(requests)".desc)
    .show(50, false)
```

```
+-----------------------------------------------------------------+-------------+
|article                                                          |sum(requests)|
+-----------------------------------------------------------------+-------------+
|Wikipédia:Accueil_principal                                      |588234       |
|Vladimir_Poutine                                                 |556280       |
|Ukraine                                                          |421460       |
|Organisation_du_traité_de_l'Atlantique_nord                      |244871       |
|Volodymyr_Zelensky                                               |244296       |
|Arme_nucléaire                                                   |138225       |
|Russie                                                           |107796       |
|Spécial:Recherche                                                |98765        |
|Guerre_du_Donbass                                                |97424        |
|Lioudmila_Poutina                                                |91800        |
|Union_des_républiques_socialistes_soviétiques                    |87341        |
|Protocole_de_Minsk                                               |73001        |
|Liste_des_armées_nationales_par_effectif                         |70336        |
|Katerina_Tikhonova                                               |66383        |
|Invasion_de_l'Ukraine_par_la_Russie_en_2022                      |65655        |
|Prolifération_nucléaire                                          |64912        |
|Catastrophe_nucléaire_de_Tchernobyl                              |63646        |
|États_membres_de_l'OTAN                                          |60646        |
|Histoire_de_l'Ukraine                                            |58507        |
|Loi_martiale                                                     |55958        |
|Liste_de_sondages_sur_l'élection_présidentielle_française_de_2022|52924        |
|Alina_Kabaeva                                                    |52896        |
|Maria_Vorontsova                                                 |51752        |
|Conflit_russo-ukrainien                                          |49750        |
|Crimée                                                           |47647        |
|Forces_armées_de_la_fédération_de_Russie                         |46415        |
|Troisième_Guerre_mondiale                                        |45786        |
|Dmitri_Medvedev                                                  |44061        |
|Forces_armées_françaises                                         |43704        |
|Raphaël_Glucksmann                                               |40786        |
|Anna_Delvey                                                      |40245        |
|Liste_des_présidents_de_la_fédération_de_Russie                  |39931        |
|Tsar_Bomba                                                       |36053        |
|Union_européenne                                                 |35099        |
|Oligarque                                                        |34386        |
|Tchétchénie                                                      |34017        |
```

Took 24 sec. Last updated by mpanahi at June 28 2023, 2:15:37 PM. (outdated)

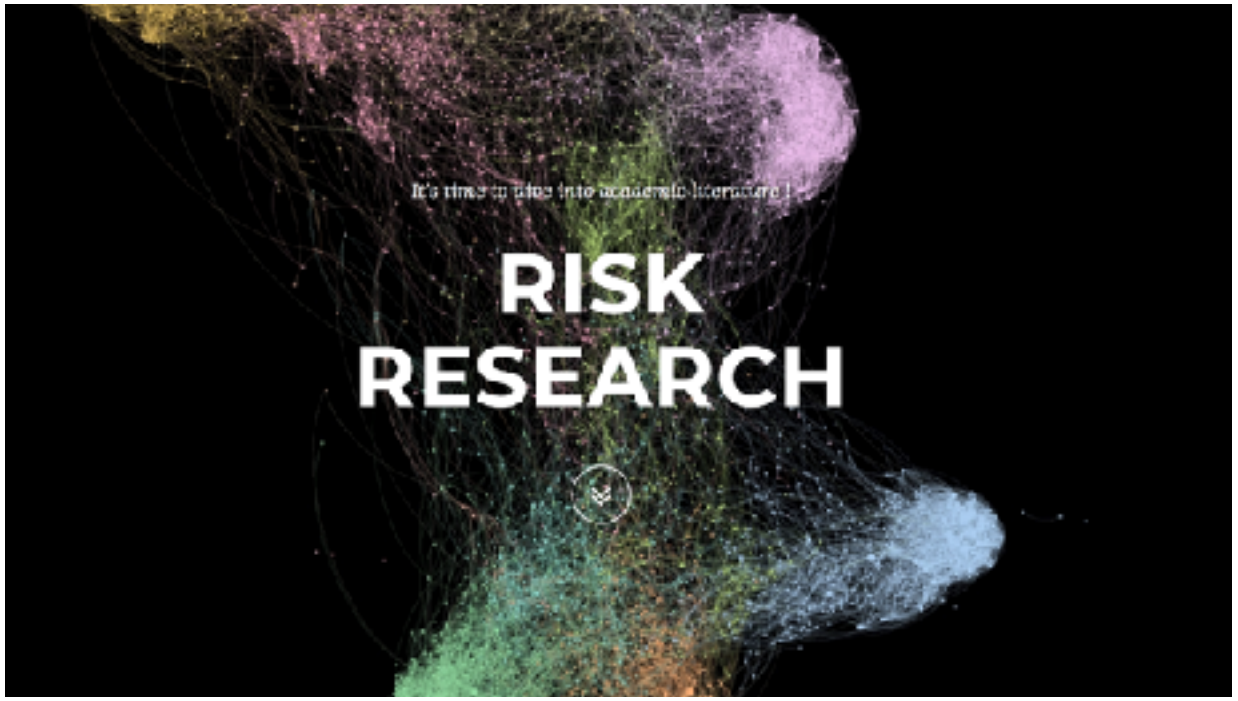


DES RÉSEAUX SOCIAUX DANS NOS URNES

LES CANDIDATS ET LEURS COMMUNAUTÉS
FACE À FACE DU SECOND TOUR
ATTENTION DES COMMUNAUTÉS

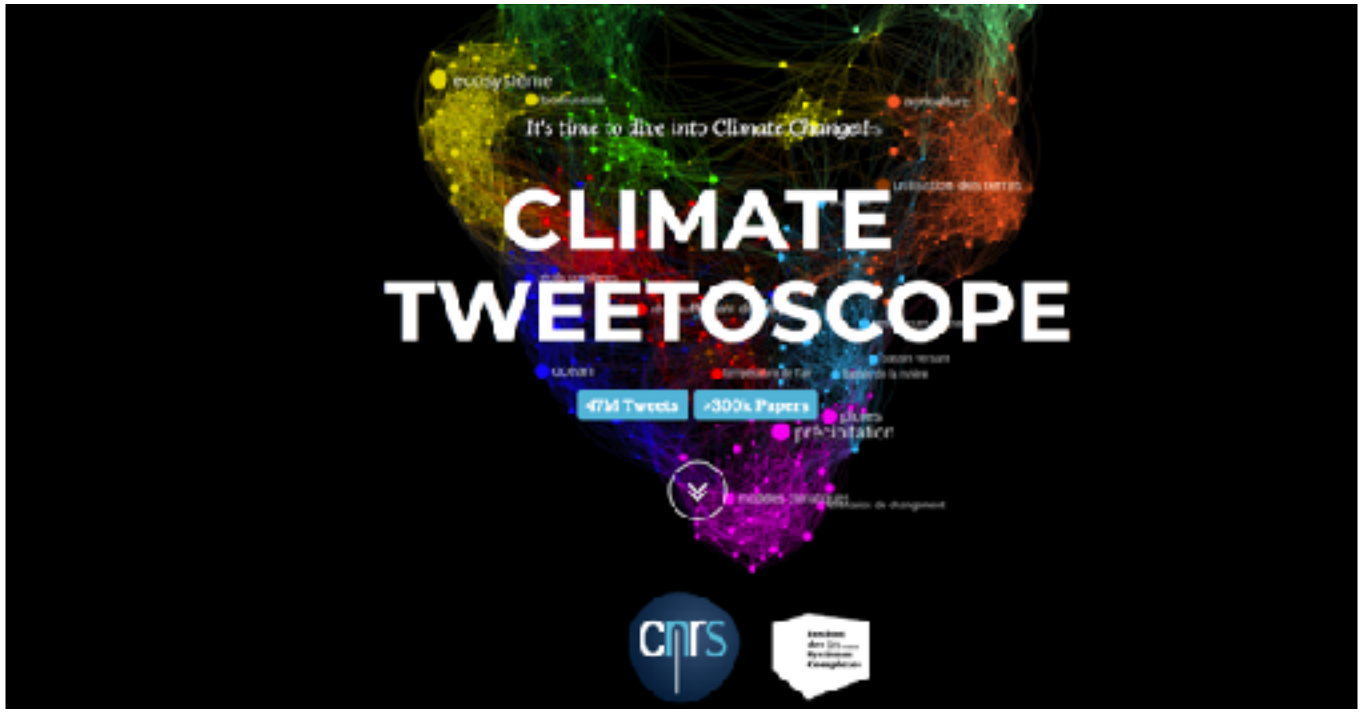
Depuis août 2016, l'ISC-PIF analyse les interactions de plus de 3000 comptes Twitter liés à la campagne présidentielle française 2017 (élus, partis, candidats) : (re)tweets, mentions, citations, etc. Une plateforme collecte et traite ainsi chaque mois plusieurs millions d'événements auxquels ont participé au total plus d'un million d'utilisateurs uniques. Une série d'analyses portant sur la structure des interactions et leur contenu est présentée dans ce dispositif, qui fait partie du Politoscope. L'objectif est de restituer dans le temps des événements perçus dans le flux continu de l'actualité. C'est une plongée inédite dans les différentes temporalités de la campagne présidentielle qui est proposée ici.

POLITOSCOPE.ORG #PRESIDENTIELLE2017



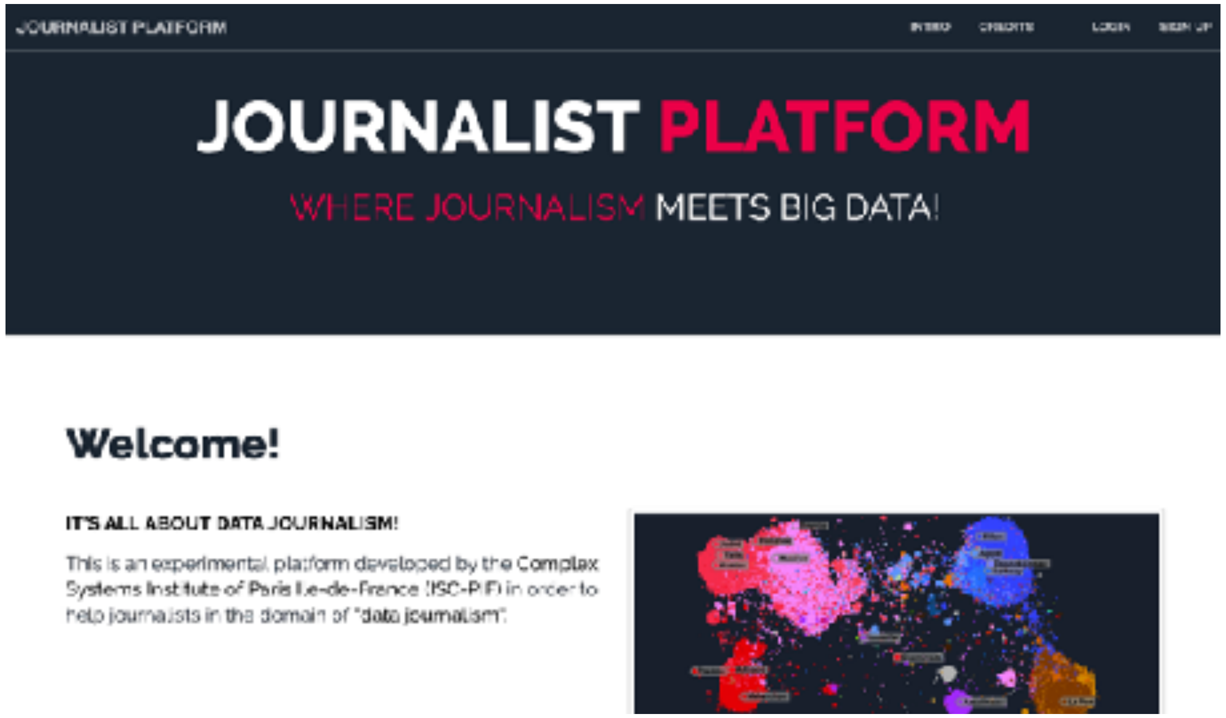
It's time to dive into academic literature !

RISK RESEARCH



It's time to dive into Climate Change!

CLIMATE TWEETOSCOPE



JOURNALIST PLATFORM

WHERE JOURNALISM MEETS BIG DATA!

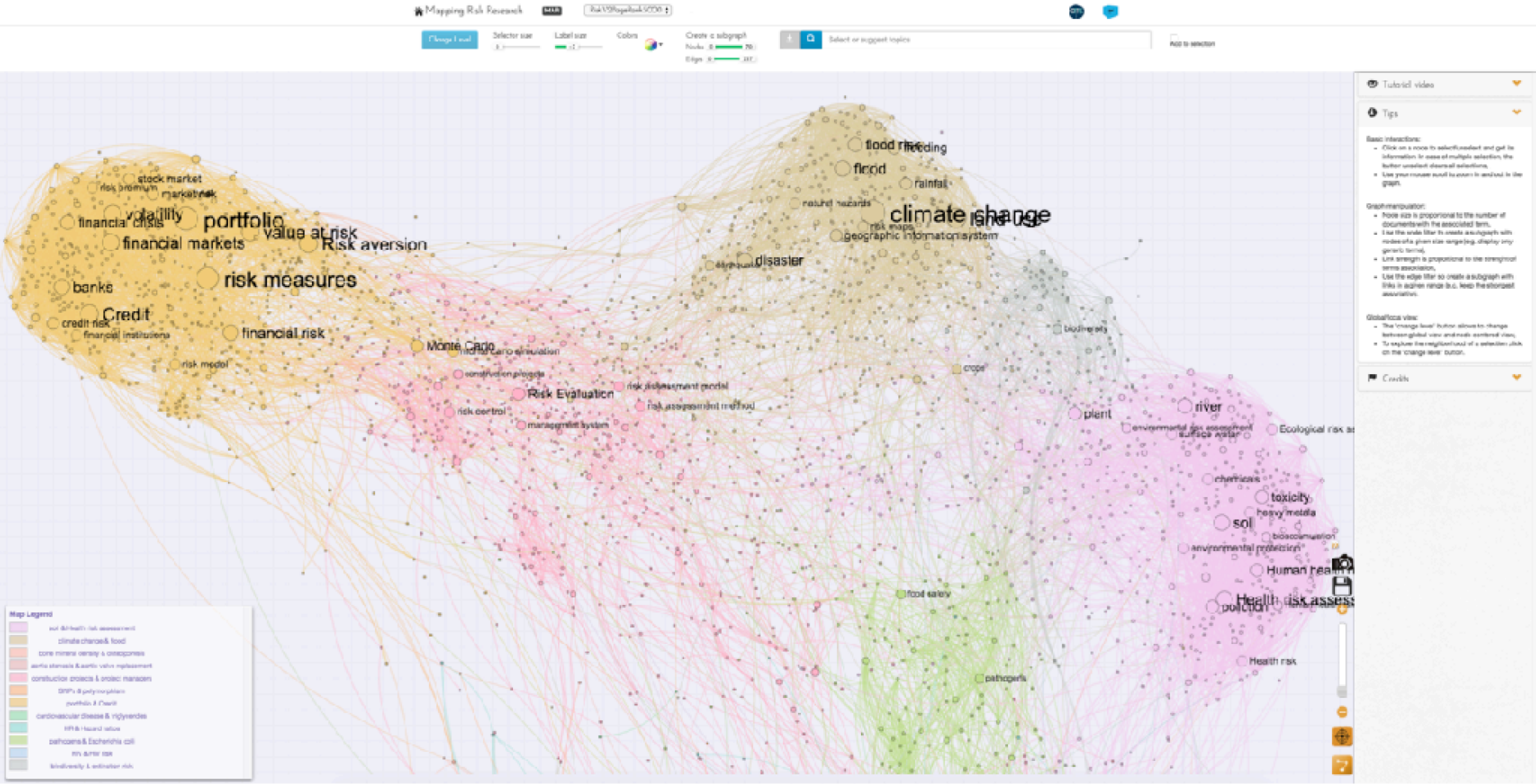
Welcome!

IT'S ALL ABOUT DATA JOURNALISM!

This is an experimental platform developed by the Complex Systems Institute of Paris Île-de-France (ISC-PIF) in order to help journalists in the domain of "data journalism".

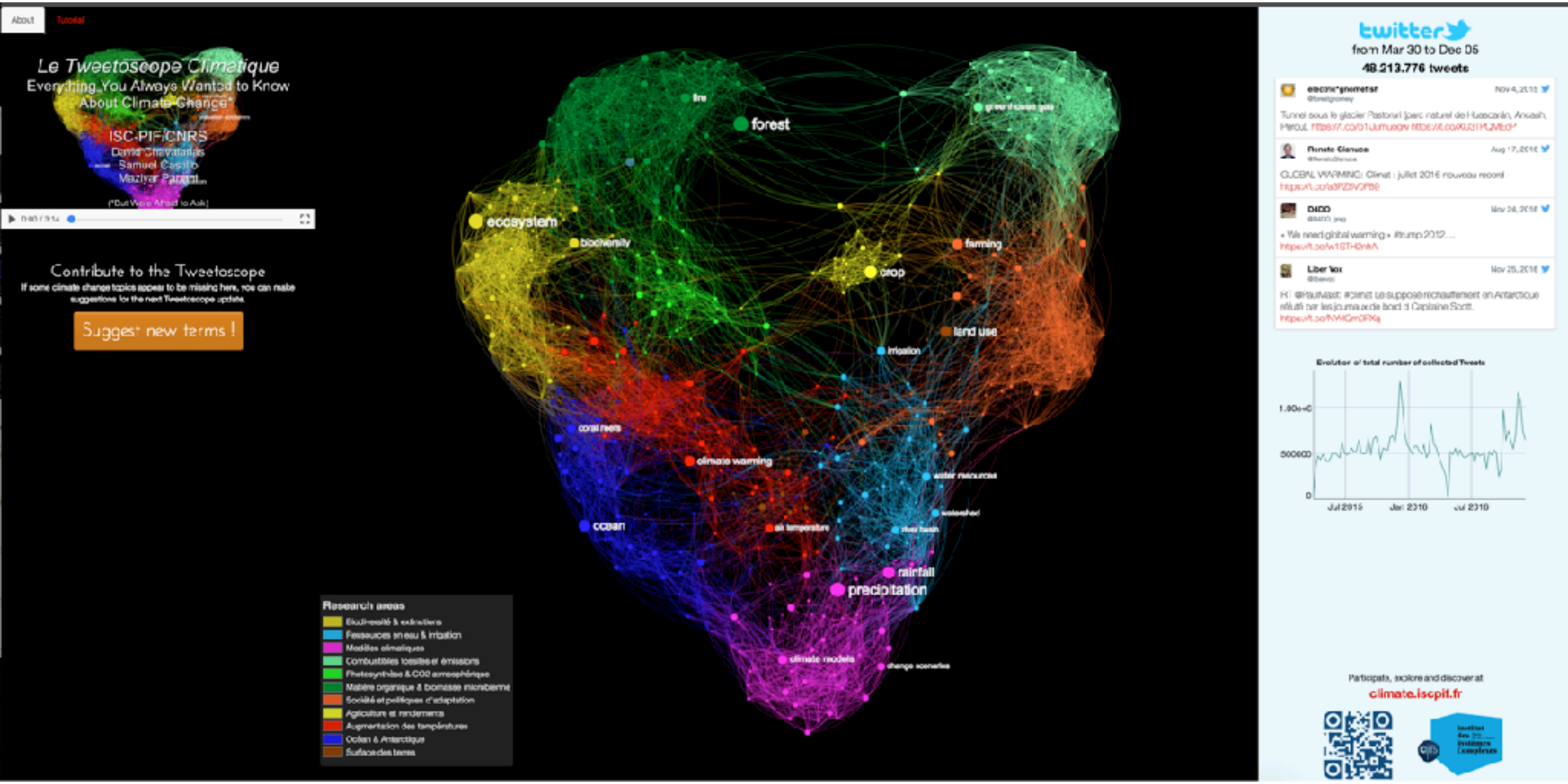
BUILT BY **MULTIVAC** PLATFORM

http://risk.iscpif.fr



INSTITUT DES **SYSTÈMES COMPLEXES** DE **PARIS** ÎLE-DE-FRANCE **(ISCPIF)**

http://tweetoscope.iscpif.fr



Le Tweetoscope Climatique
Everything You Always Wanted to Know About Climate Change*
ISC-PIF/CNRS
David Chavalarias
Samuel Castillo
Maziyar Panahi
(*But Were Afraid to Ask)

Contribute to the Tweetoscope
If some climate change topics appear to be missing here, you can make suggestions for the next Tweetoscope update.

twitter
from Mar 30 to Dec 05
48.213.776 tweets

Tunnel sous le glacier Pastoruri (parc naturel de Huascarán, Ancash, Pérou)

GLOBAL WARMING: Climat : juillet 2016 nouveau record

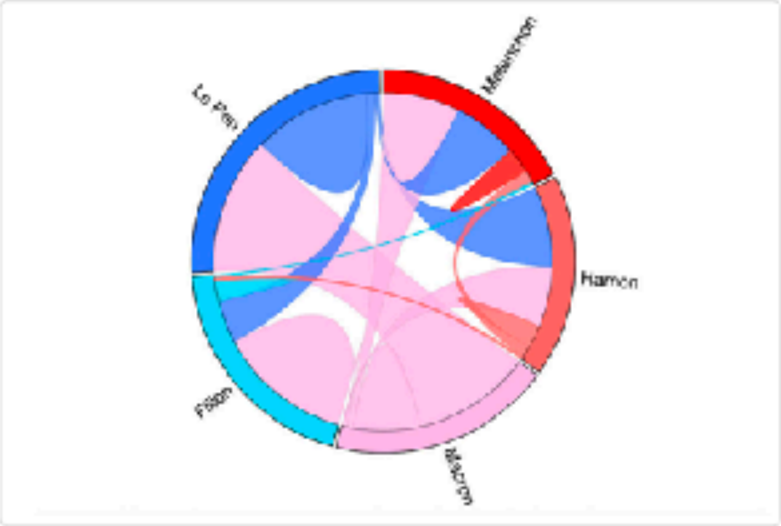
« We need global warming » #trump2012…

RT #climat Le supposé réchauffement en Antarctique réfuté par les journaux de bord di Capitaine Scott.

Participate, explore and discover at
climate.iscpif.fr

INSTITUT DES SYSTÈMES COMPLEXES DE PARIS ÎLE-DE-FRANCE (ISCPIF)

DES RÉSEAUX SOCIAUX DANS NOS URNES



LES CANDIDATS ET LEURS COMMUNAUTÉS

FACE À FACE DU SECOND TOUR

ATTENTION DES COMMUNAUTÉS

Depuis août 2016, l'ISC-PIF analyse les interactions de plus de 3000 comptes Twitter liés à la campagne présidentielle française 2017 (élus, partis, candidats) : (re)tweets, mentions, citations, etc. Une plateforme collecte et traite ainsi chaque mois plusieurs millions d'événements auxquels ont participé au total plus d'un million d'utilisateurs uniques. Une série d'analyses portant sur la structure des interactions et leur contenu est présentée dans ce dispositif, qui fait partie du *Politoscope*. L'objectif est de resituer dans le temps des événements perçus dans le flux continu de l'actualité. C'est une plongée inédite dans les différentes temporalités de la campagne présidentielle qui est proposée ici.

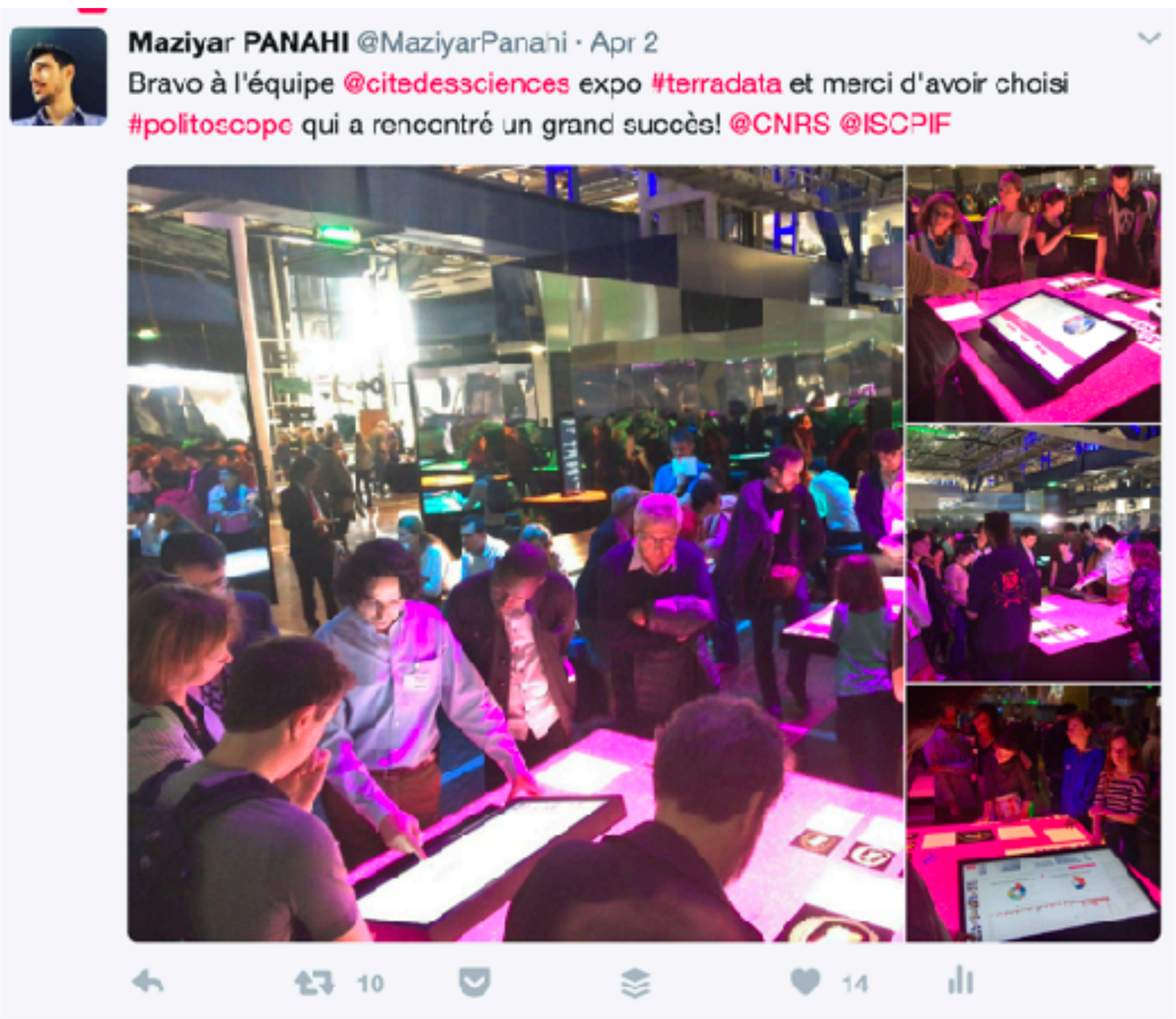
MÉTHODOLOGIE

CRÉDITS

POLITOSCOPE.ORG #PRESIDENTIELLE2017

http://politoscope.org

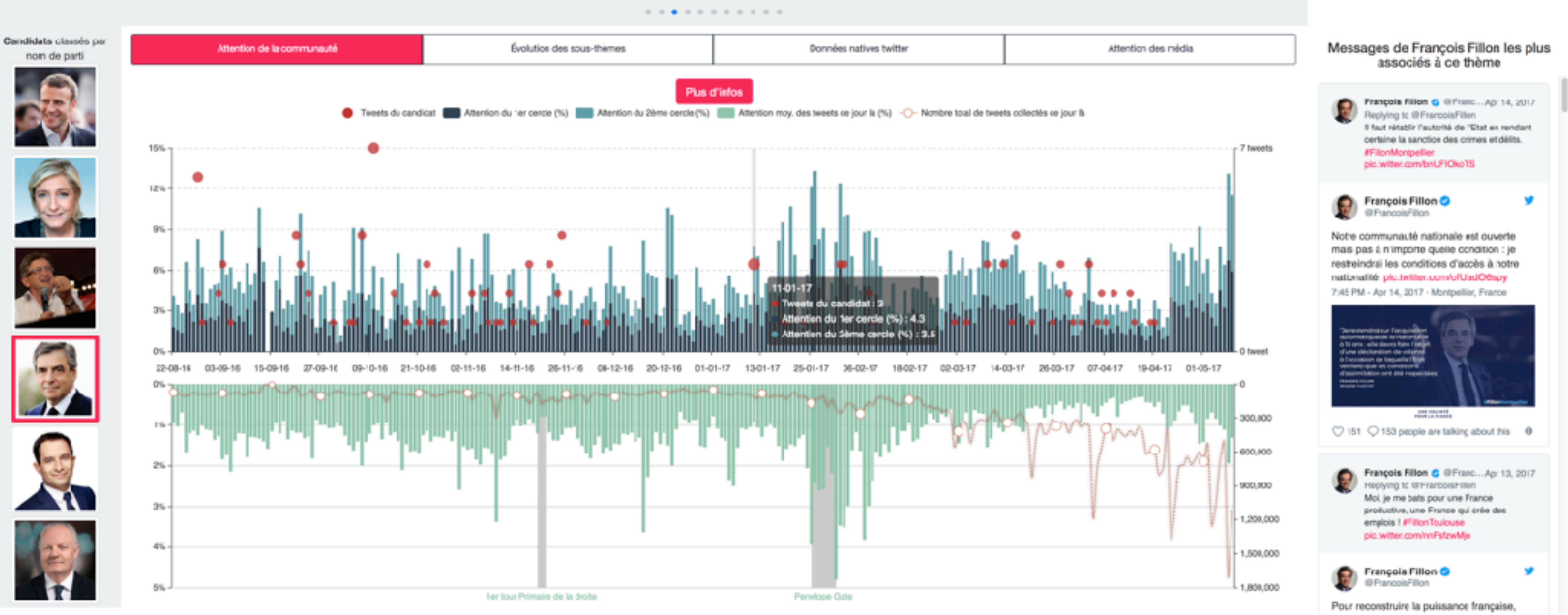
https://politoscope.org



Maziyar PANAHI @MaziyarPanahi · Apr 2

Bravo à l'équipe @citedessciences expo #terradata et merci d'avoir choisi #politoscope qui a rencontré un grand succès! @CNRS @ISCPIF

cité des sciences & de l'industrie
TERRA DATA
NOS VIES À L'ÈRE DU NUMÉRIQUE
EXPOSITION DU 04/04/17 AU 07/01/18
#TerraData
CNIL
SAFRAN
Qwant
20 minutes
Alternatives Économiques
SCIENCE
franceinfo:



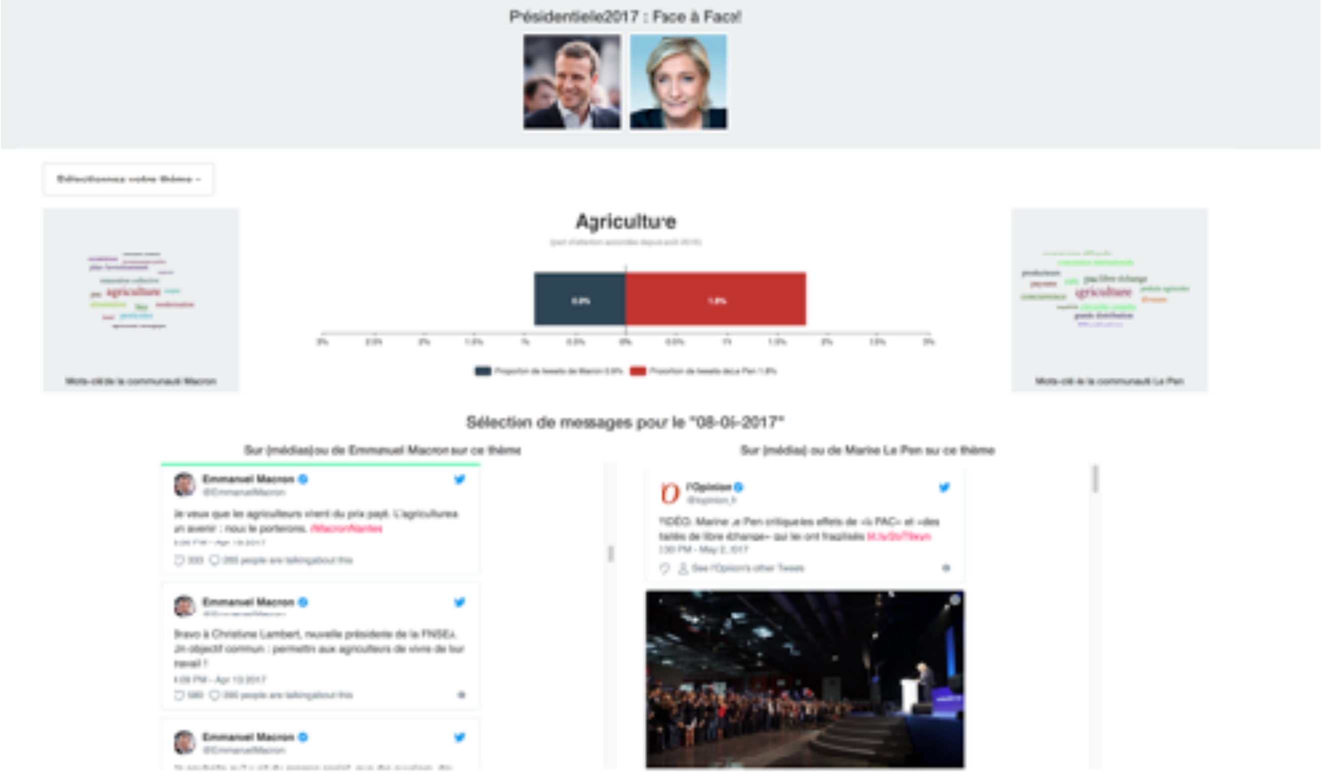
Messages de François Fillon les plus associés à ce thème

François Fillon @Franc... Apr 14, 2017
Replying to @FrancoisFillon
Il faut rétablir l'autorité de l'État en rendant certaine la sanction des crimes et délits. #FillonMontpellier pic.twitter.com/bnUFIOko1S

François Fillon @FrancoisFillon
Notre communauté nationale est ouverte mais pas à n'importe quelle condition : je restreindrai les conditions d'accès à notre nationalité. pic.twitter.com/ufUaJO6spy
7:45 PM - Apr 14, 2017 · Montpellier, France
151 · 153 people are talking about this

François Fillon @Franc... Apr 13, 2017
Replying to @FrancoisFillon
Moi, je me bats pour une France productive, une France qui crée des emplois ! #FillonToulouse pic.twitter.com/nnFsfzwMje

François Fillon @FrancoisFillon
Pour reconstruire la puissance française,


Présidentielle2017 : Face à Face!
Sélectionnez votre thème
Agriculture
agriculture
Mots-clé de la communauté Macron
Mots-clé de la communauté Le Pen

Sélection de messages pour le "08-05-2017"

Sur (médias) ou de Emmanuel Macron sur ce thème

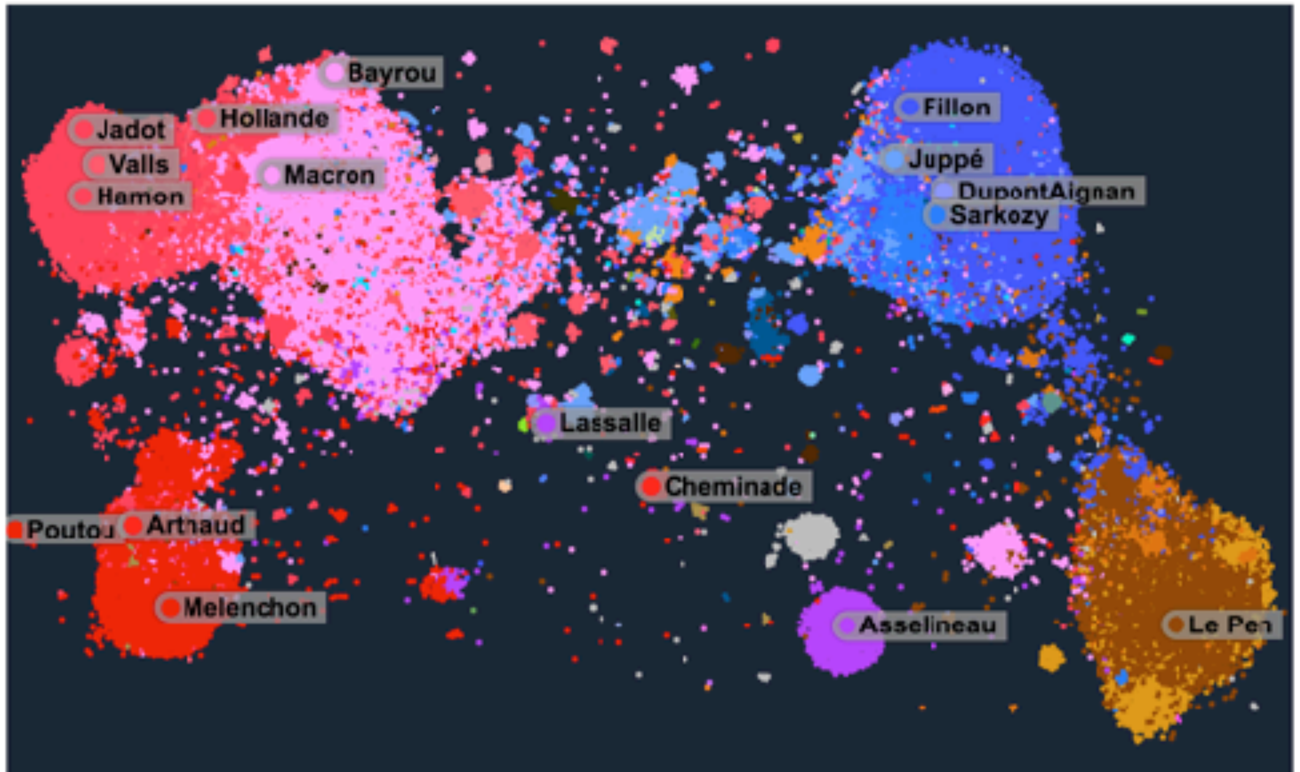
Emmanuel Macron @EmmanuelMacron
Je veux que les agriculteurs vivent du prix payé. L'agriculture a un avenir : nous le porterons. #MacronNantes

Emmanuel Macron @EmmanuelMacron
Bravo à Christiane Lambert, nouvelle présidente de la FNSEA. Un objectif commun : permettre aux agriculteurs de vivre de leur travail !

Emmanuel Macron @EmmanuelMacron

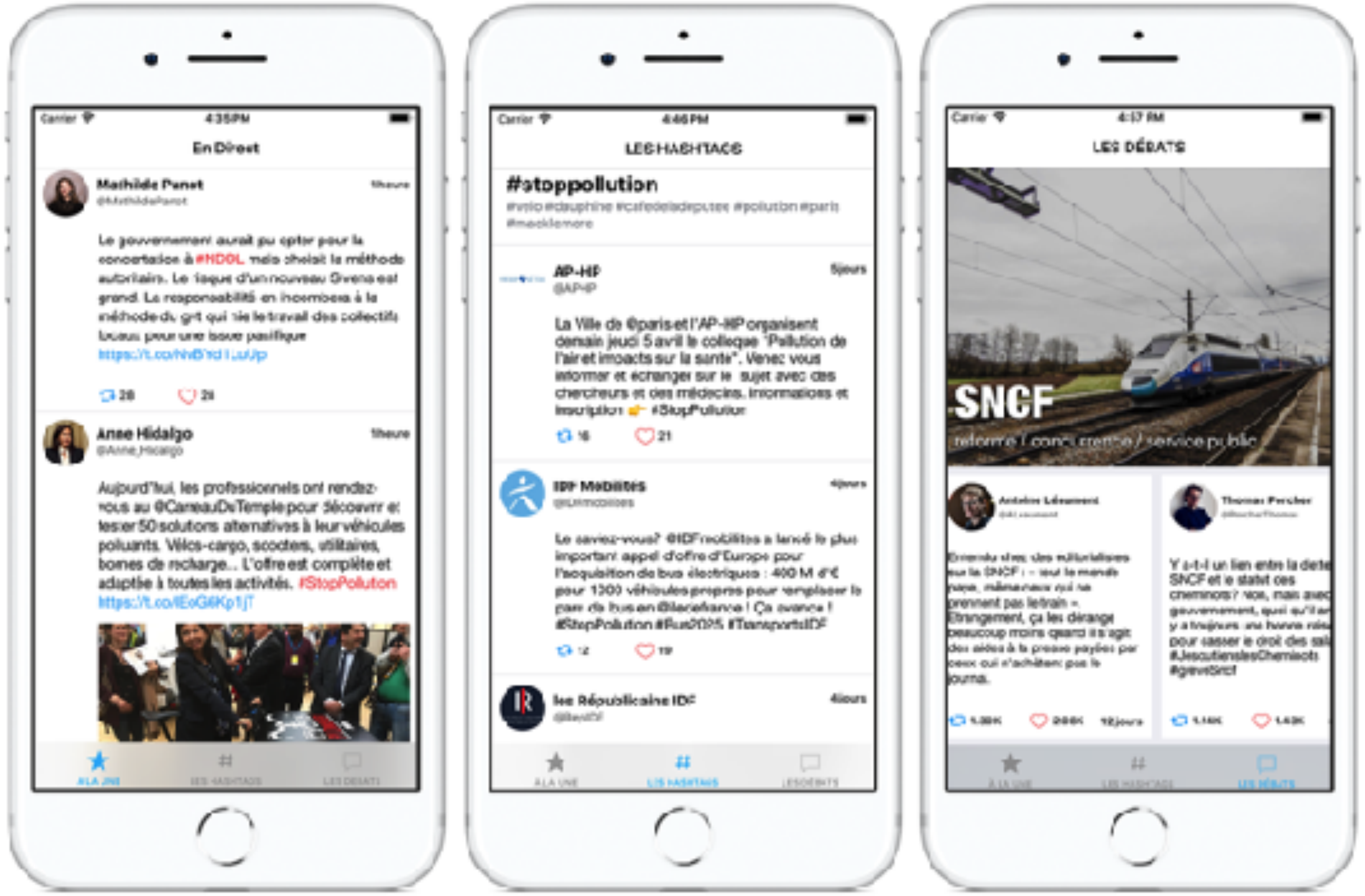
Sur (média) ou de Marine Le Pen sur ce thème

l'Opinion @lopinion_fr
VIDÉO. Marine Le Pen critique les effets de « la PAC » et « des traités de libre échange » qui les ont fragilisés
1:30 PM - May 2, 2017
See l'Opinion's other Tweets



par nom de parti
Sarkozy
Sarkozy
Sarkozy
Sarkozy, Fillon
Sarkozy, Fillon
Sarkozy, Fillon
Sarkozy, Fillon
Sarkozy, Fillon
Fillon
Fillon
Fillon
Fillon
Juppe, Fillon
Lassalle
Dupont-Aignan
9 janvier
25 janvier
Le rassemblement autour de Fillon écorné par la bataille des investitures
Bayrou
NKM
Le Maire

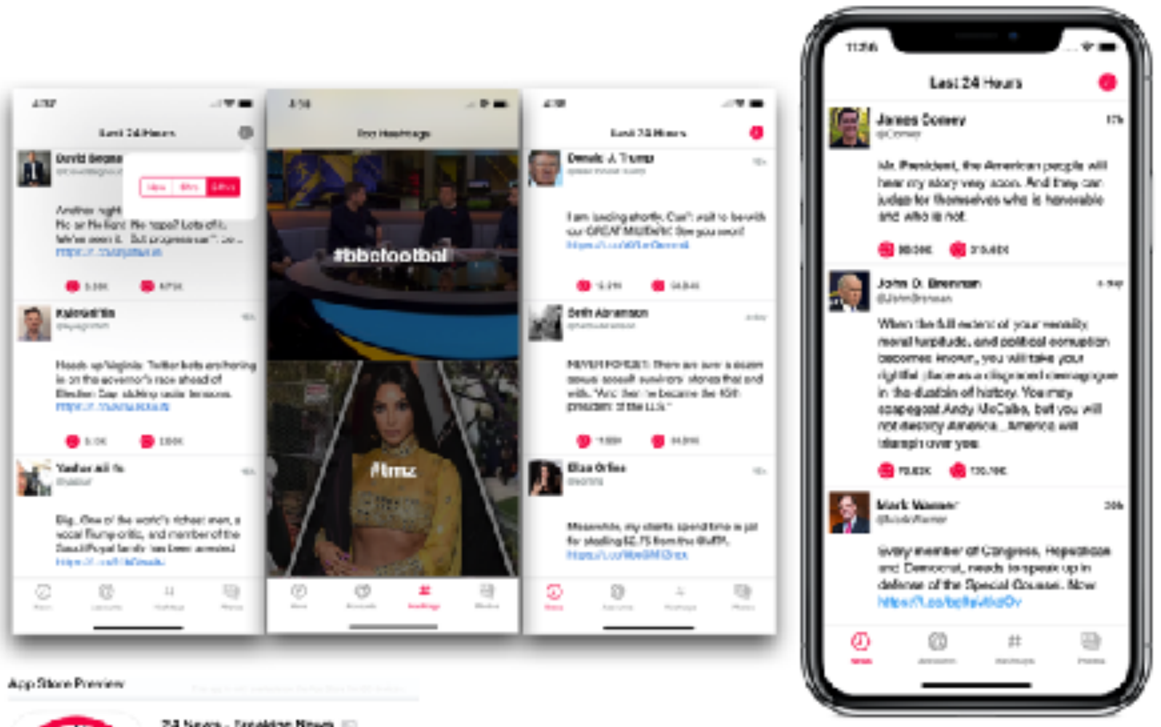
MOBILE APPLICATIONS



POLITOSCOPE: Politique en France

24 News - Real-time breaking news and headlines

24 NEWS: Real-time News and Highlights



App Store Preview

24 News - Breaking News

24 NEWS: Real-time News and Highlights

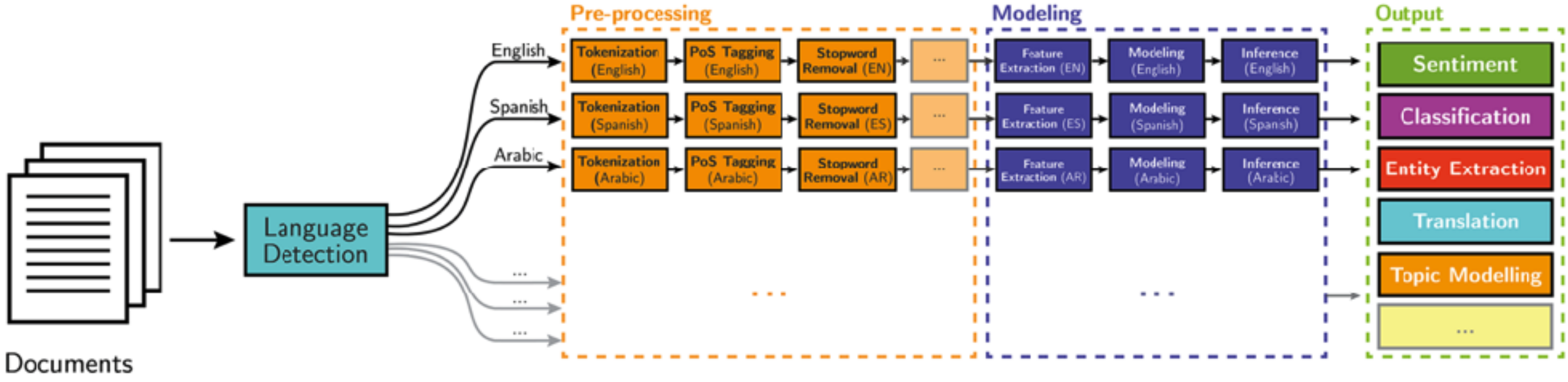
MULTIVAC PLATFORM
The Platform of Platforms

BUILT BY **MULTIVAC** PLATFORM

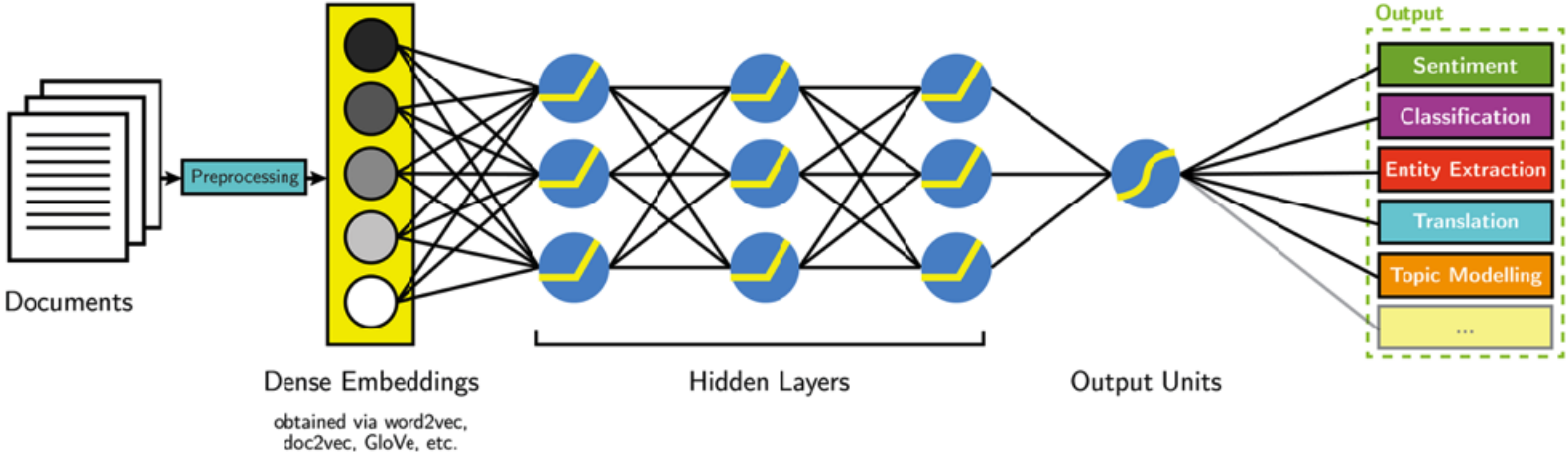
Le grand
débat national

Comment les nouvelles technologies changent les rapports que les individus entretiennent avec les productions numériques collectives ?

Classical NLP



Deep Learning-based NLP



122,309
Unique Users

251,244
Proposals

2,192,163
Answers

February 17th

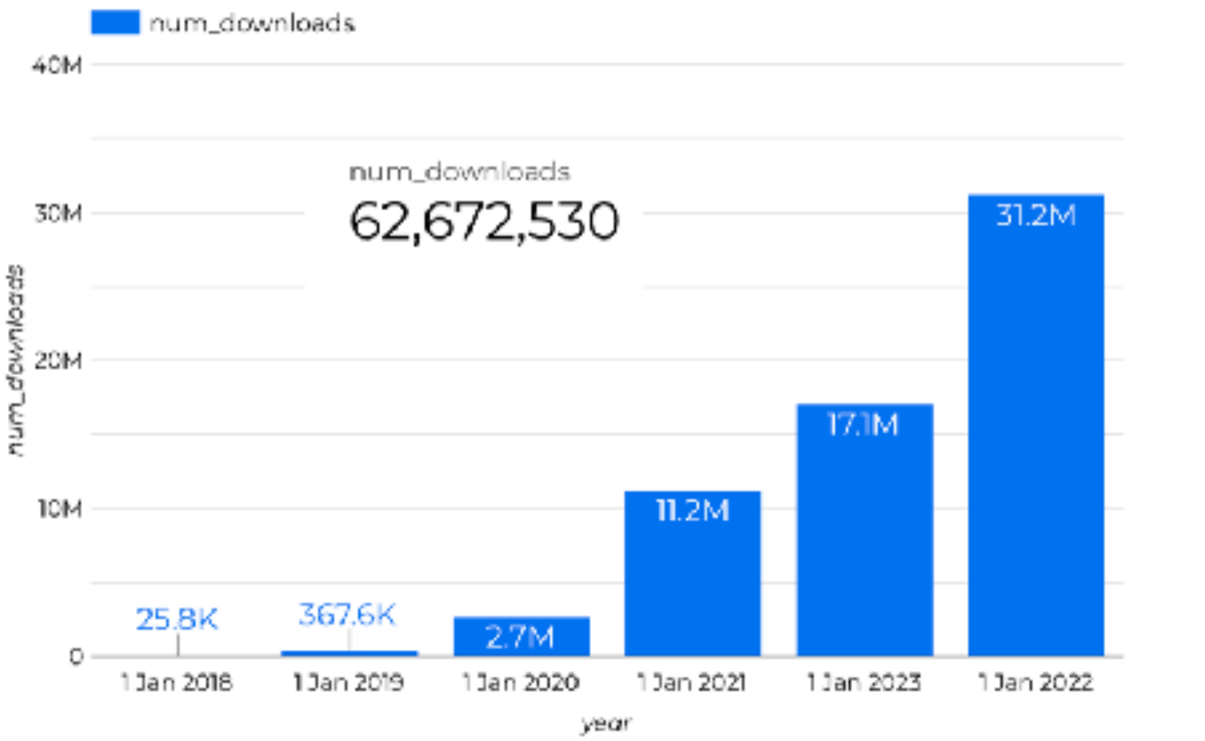
840K
Unique Words

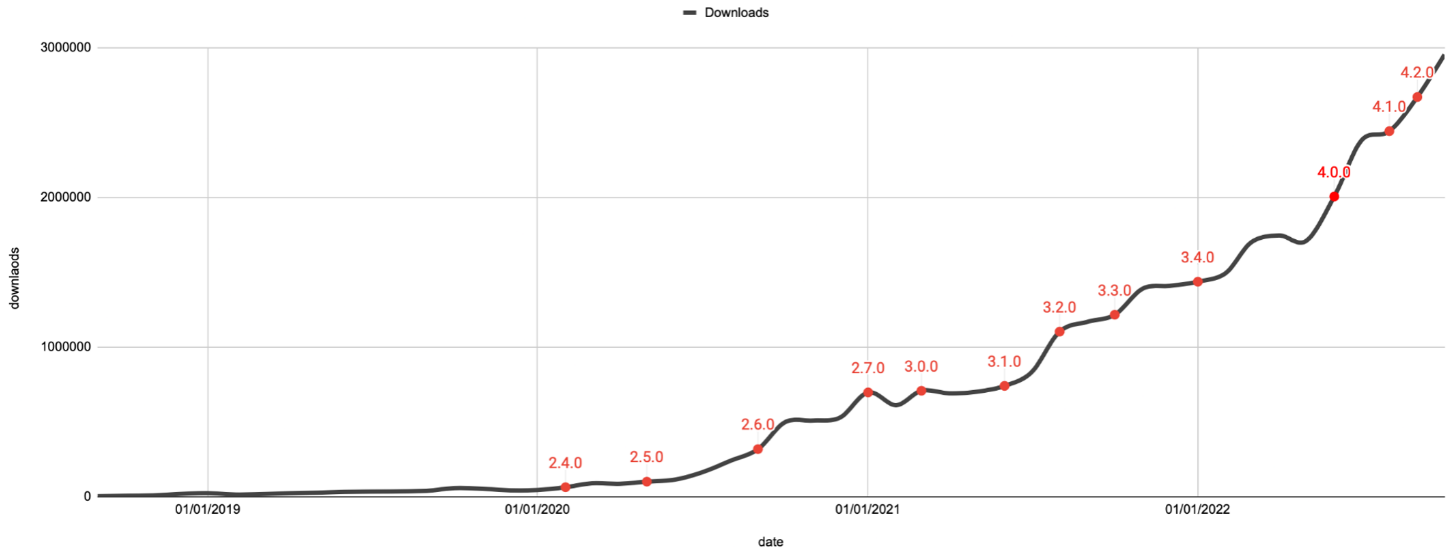
66M
Total Words

+1 Million
Answers > **100** Characters

+3.5 Million
Unique Phrases

Introducing Spark NLP





Spark NLP is an open-source natural language processing library, built on top of **Apache Spark** and **Spark ML**. (first release: July 2017)

- A single unified solution for all your NLP needs
- Take advantage of transfer learning and implementing the latest and greatest **SOTA** algorithms and models in NLP research
- The most widely used NLP library in industry (5 yrs in a row)

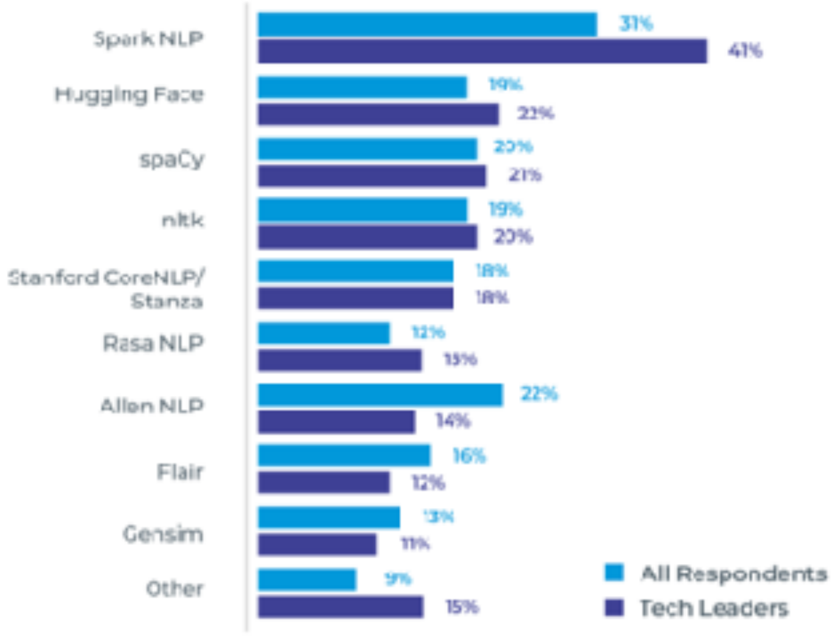
The most scalable, accurate and fastest library in NLP history

- 111 total releases, every two weeks for the past 5 years

Spark NLP in Industry

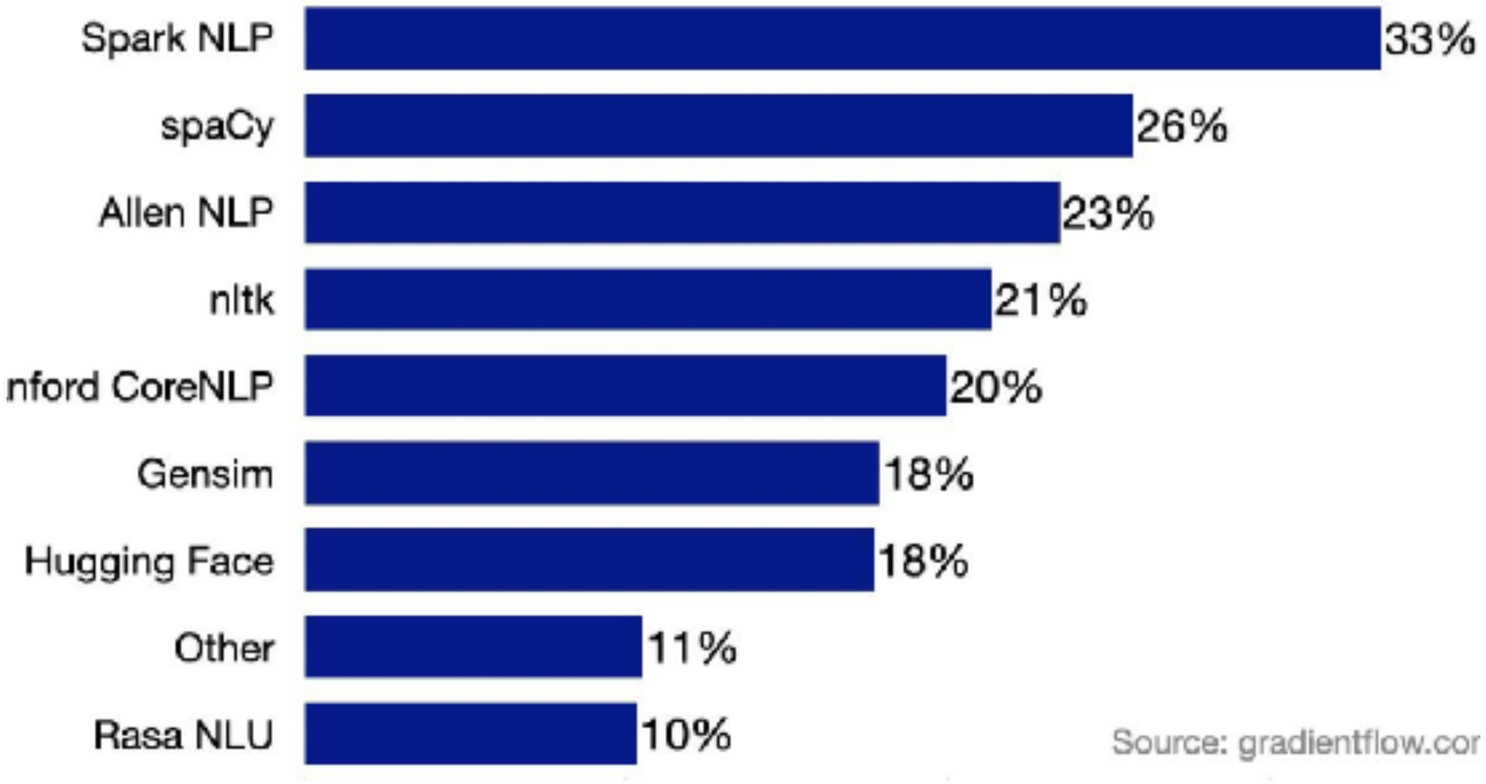
The most widely used NLP library in the enterprise

Which NLP Libraries does your organization use in production?



Gradient Flow NLP Survey, 2021.

Which NLP libraries does your organization use?



Source: gradientflow.cor

NLP Industry Survey by Gradient Flow,

an independent data science research & insights company, September 2020/2021

Entity Recognition
I love Lucy PERSON

Information Extraction
They met Last week DATE -> 29-04-2020

Spelling & Grammar
abc She become the first... -> She became the first

Text Classification

Image Classification

Translation
[je t'aime -> i love you]

Summarization

Question Answering
Q&A

Emotion Detection

Automatic Speech Recognition

Split Text

- Sentence Detector
- Tokenizer
- Normalizer
- nGram Generator
- Word Segmentation

Clean Text

- Spell Checker
- Grammar Checker
- Writing Style Checker
- Stopword Cleaner
- Summarization

Understand Grammar

- Stemmer
- Lemmatizer
- Part of Speech Tagger
- Dependency Parser
- Translation

Find in Text

- Text Matcher
- Regex Matcher
- Date Matcher
- Chunker
- Question Answering

17,000+

~~12,000+~~

Pre-trained Pipelines, Models & Transformers

BERT ELMO TAPAS
ALBERT DeBERTa USE
Longformer ELECTRA
T5 NMT ViT
DistilBERT RoBERTa
XLM-RoBERTa
Wav2Vec2 XLNet

250+

Languages

Trainable & Tunable

Scalable to a Cluster
APACHE Spark ML Pipelines

Fast Inference
LightPipeline

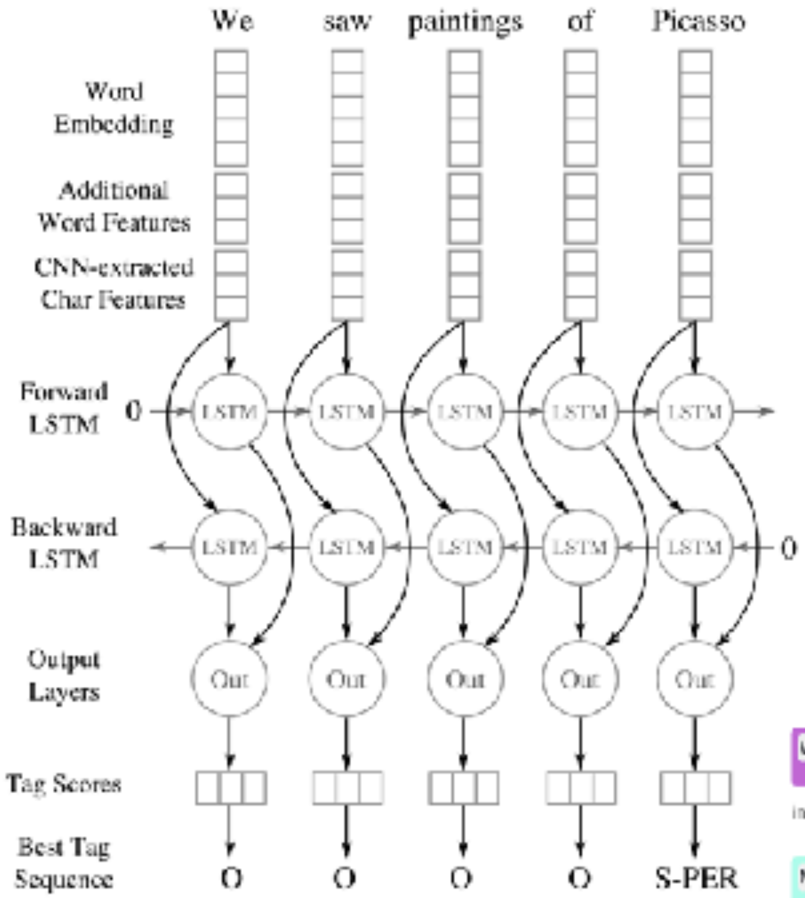
Hardware Optimized
intel NVIDIA

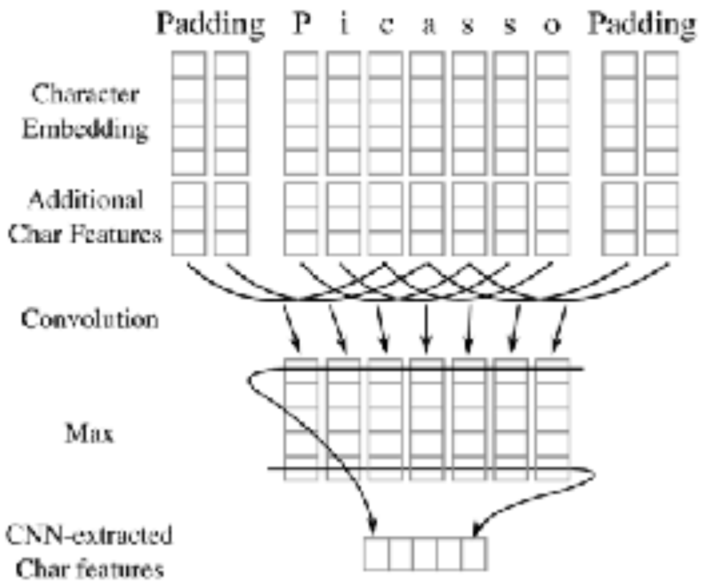
Community
NLP SUMMIT

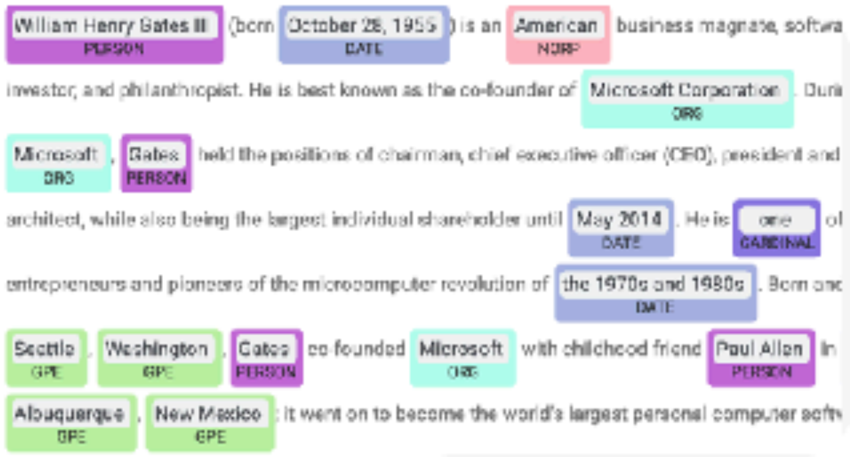
Accuracy: State-of-the-art Models

Named Entity Recognition

- "State of the art" means the best peer-reviewed **academic results**
- For example: Best **F1 score** on **CoNLL-2003 NER** benchmark for a system in production
- Spark NLP uses **Bi-LSTM + Char-CNN + CRF + Word Embeddings**






William Henry Gates III PERSON (born October 28, 1955 DATE) is an American NORP business magnate, softwa

investor, and philanthropist. He is best known as the co-founder of Microsoft Corporation ORG . Duri

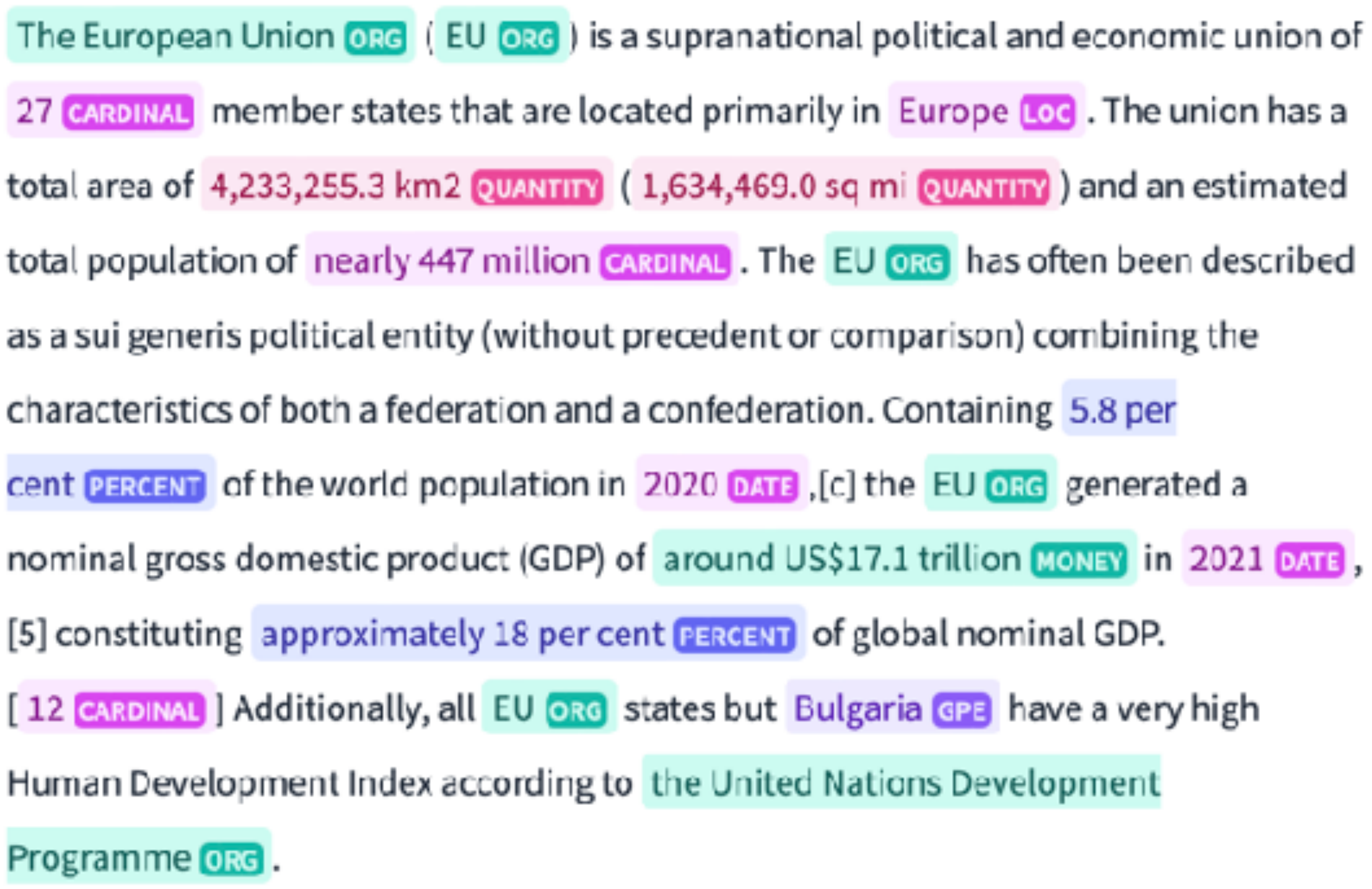
Microsoft ORG , Gates PERSON held the positions of chairman, chief executive officer (CEO), president and

architect, while also being the largest individual shareholder until May 2014 DATE . He is one CARDINAL of

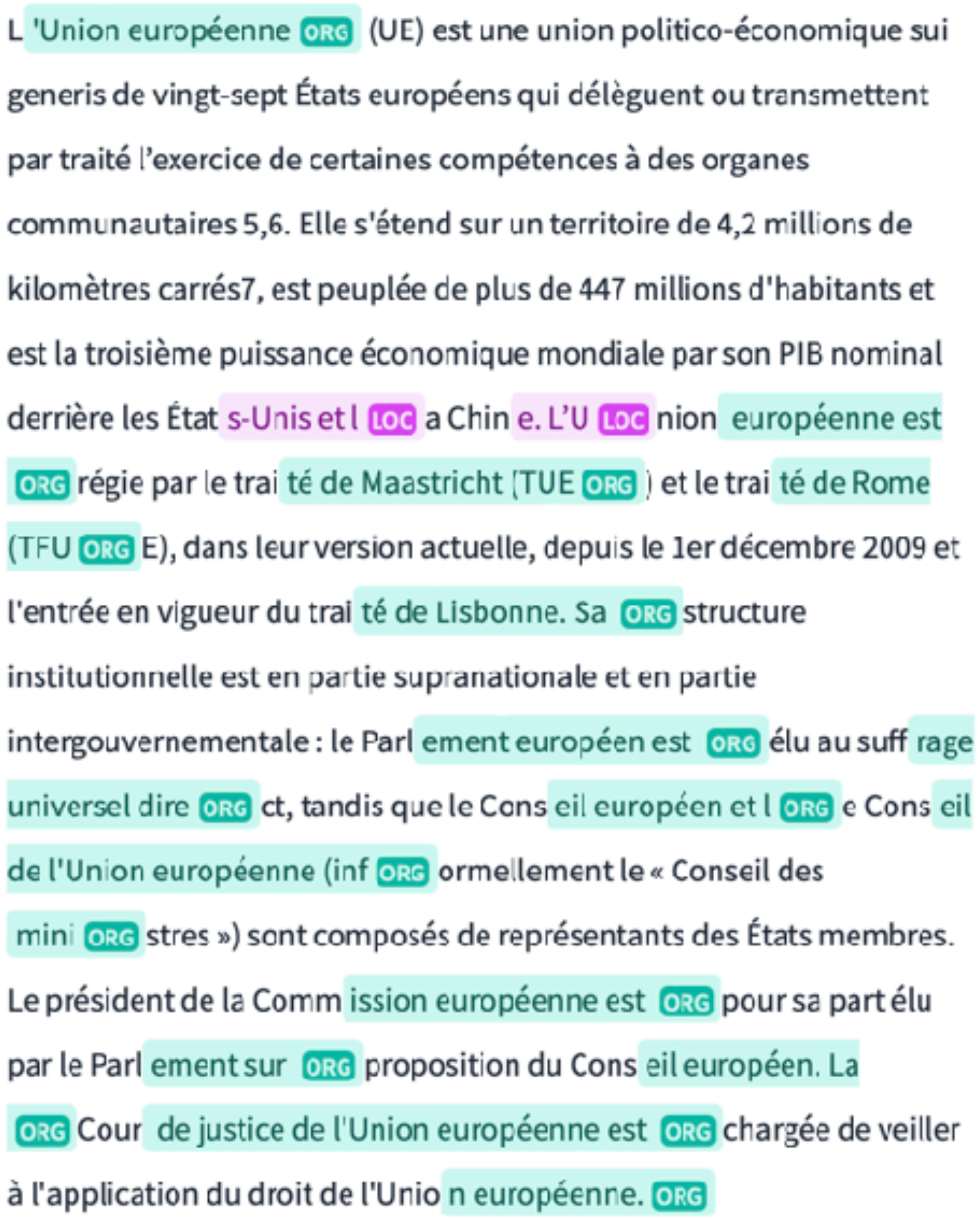
entrepreneurs and pioneers of the microcomputer revolution of the 1970s and 1980s DATE . Born and

Seattle GPE , Washington GPE , Gates PERSON co-founded Microsoft ORG with childhood friend Paul Allen PERSON in

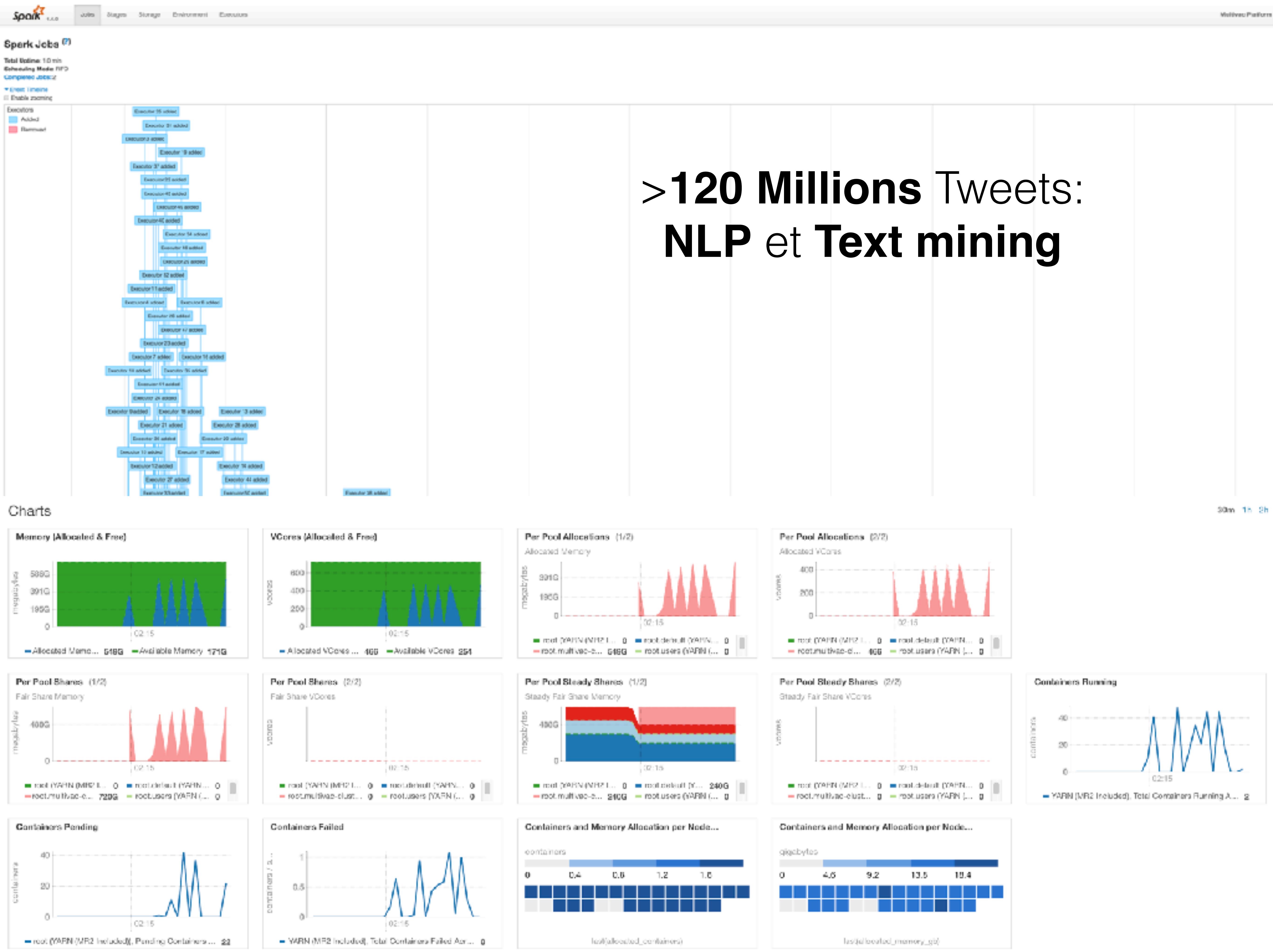
Albuquerque GPE , New Mexico GPE ; it went on to become the world's largest personal computer softw



The European Union ORG (EU ORG) is a supranational political and economic union of 27 CARDINAL member states that are located primarily in Europe LOC . The union has a total area of 4,233,255.3 km2 QUANTITY (1,634,469.0 sq mi QUANTITY) and an estimated total population of nearly 447 million CARDINAL . The EU ORG has often been described as a sui generis political entity (without precedent or comparison) combining the characteristics of both a federation and a confederation. Containing 5.8 per cent PERCENT of the world population in 2020 DATE ,[c] the EU ORG generated a nominal gross domestic product (GDP) of around US$17.1 trillion MONEY in 2021 DATE , [5] constituting approximately 18 per cent PERCENT of global nominal GDP. [12 CARDINAL] Additionally, all EU ORG states but Bulgaria GPE have a very high Human Development Index according to the United Nations Development Programme ORG .

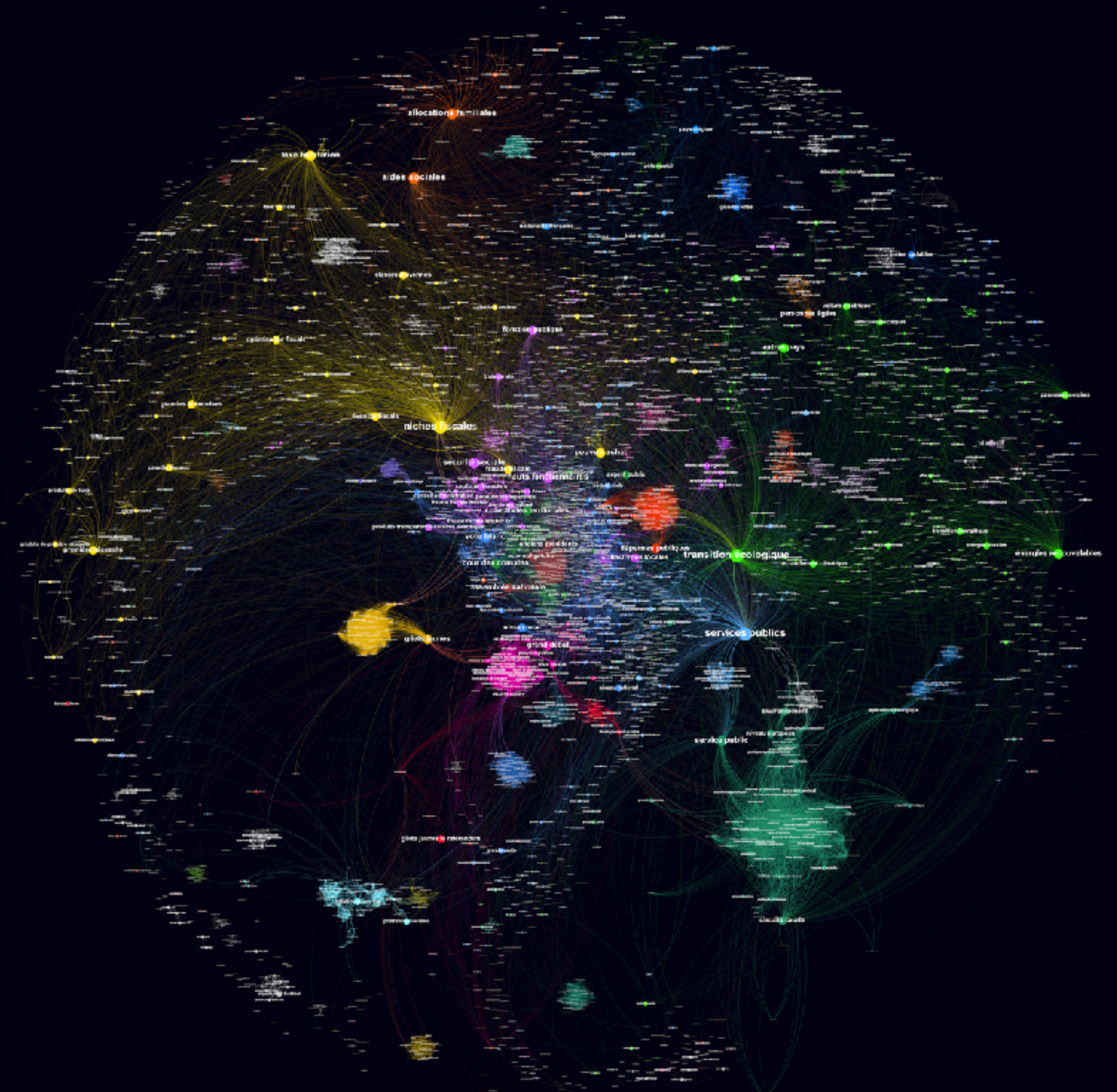


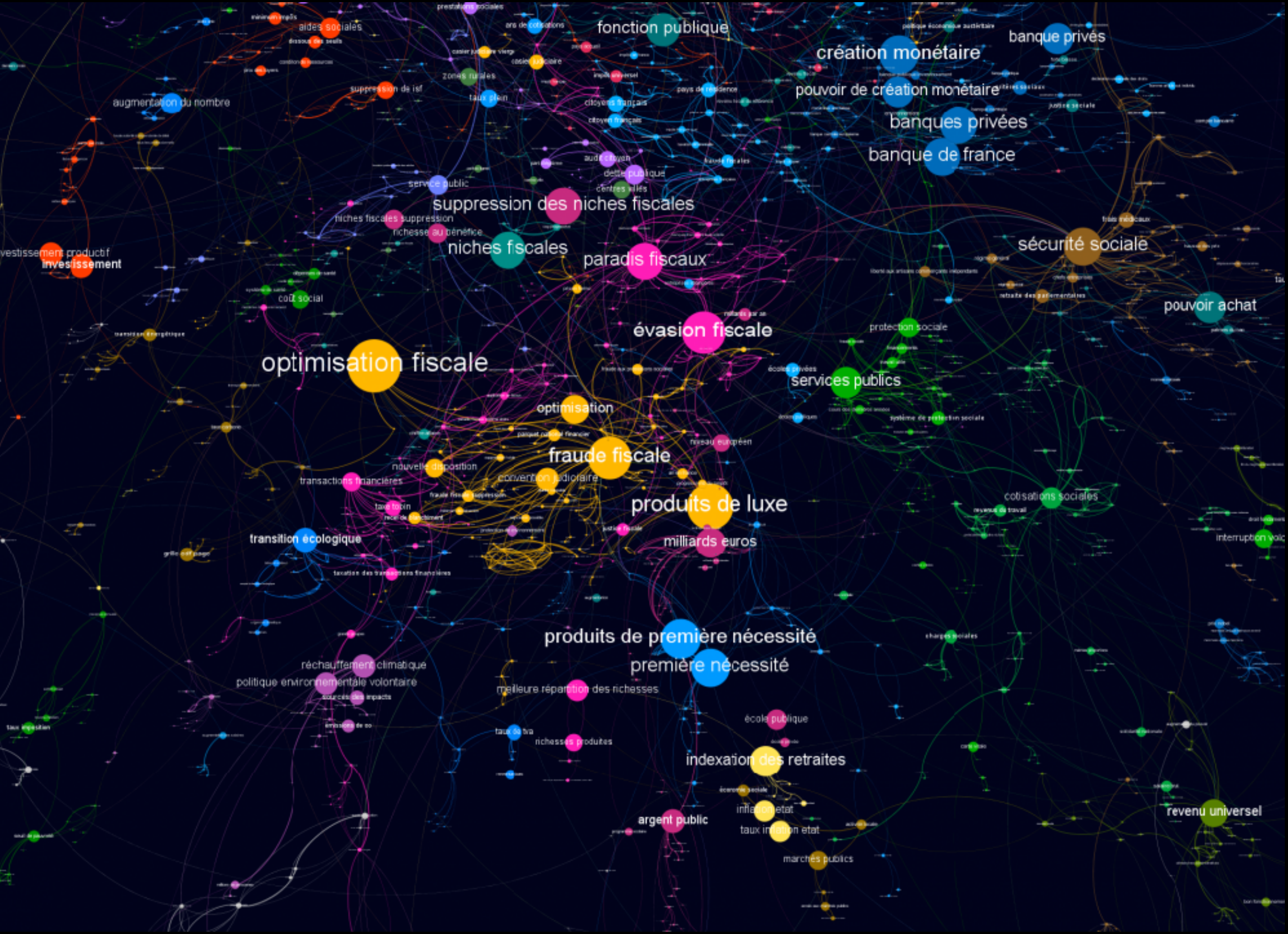
L 'Union européenne ORG (UE) est une union politico-économique sui generis de vingt-sept États européens qui délèguent ou transmettent par traité l'exercice de certaines compétences à des organes communautaires 5,6. Elle s'étend sur un territoire de 4,2 millions de kilomètres carrés7, est peuplée de plus de 447 millions d'habitants et est la troisième puissance économique mondiale par son PIB nominal derrière les État s-Unis et l LOC a Chin e. L'U LOC nion européenne est ORG régie par le trai té de Maastricht (TUE ORG) et le trai té de Rome (TFU ORG E), dans leur version actuelle, depuis le 1er décembre 2009 et l'entrée en vigueur du trai té de Lisbonne. Sa ORG structure institutionnelle est en partie supranationale et en partie intergouvernementale : le Parl ement européen est ORG élu au suff rage universel dire ORG ct, tandis que le Cons eil européen et l ORG e Cons eil de l'Union européenne (inf ORG ormellement le « Conseil des mini ORG stres ») sont composés de représentants des États membres. Le président de la Comm ission européenne est ORG pour sa part élu par le Parl ement sur ORG proposition du Cons eil européen. La ORG Cour de justice de l'Union européenne est ORG chargée de veiller à l'application du droit de l'Unio n européenne. ORG

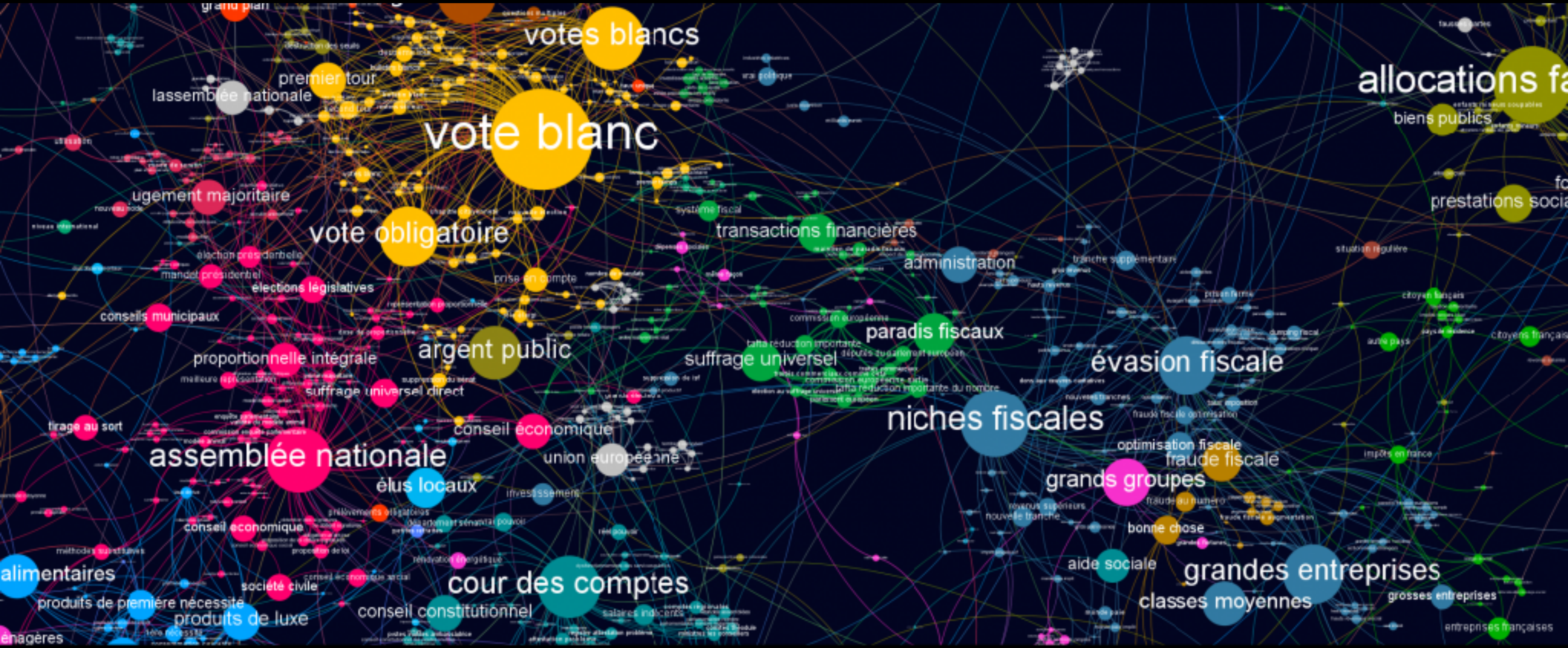

Spark
Spark Jobs
Executors
Added
Removed
Charts
Memory (Allocated & Free)
VCores (Allocated & Free)
Per Pool Allocations (1/2)
Per Pool Allocations (2/2)
Per Pool Shares (1/2)
Per Pool Shares (2/2)
Per Pool Steady Shares (1/2)
Per Pool Steady Shares (2/2)
Containers Running
Containers Pending
Containers Failed
Containers and Memory Allocation per Node...
Containers and Memory Allocation per Node...

>**120 Millions** Tweets:

NLP et **Text mining**

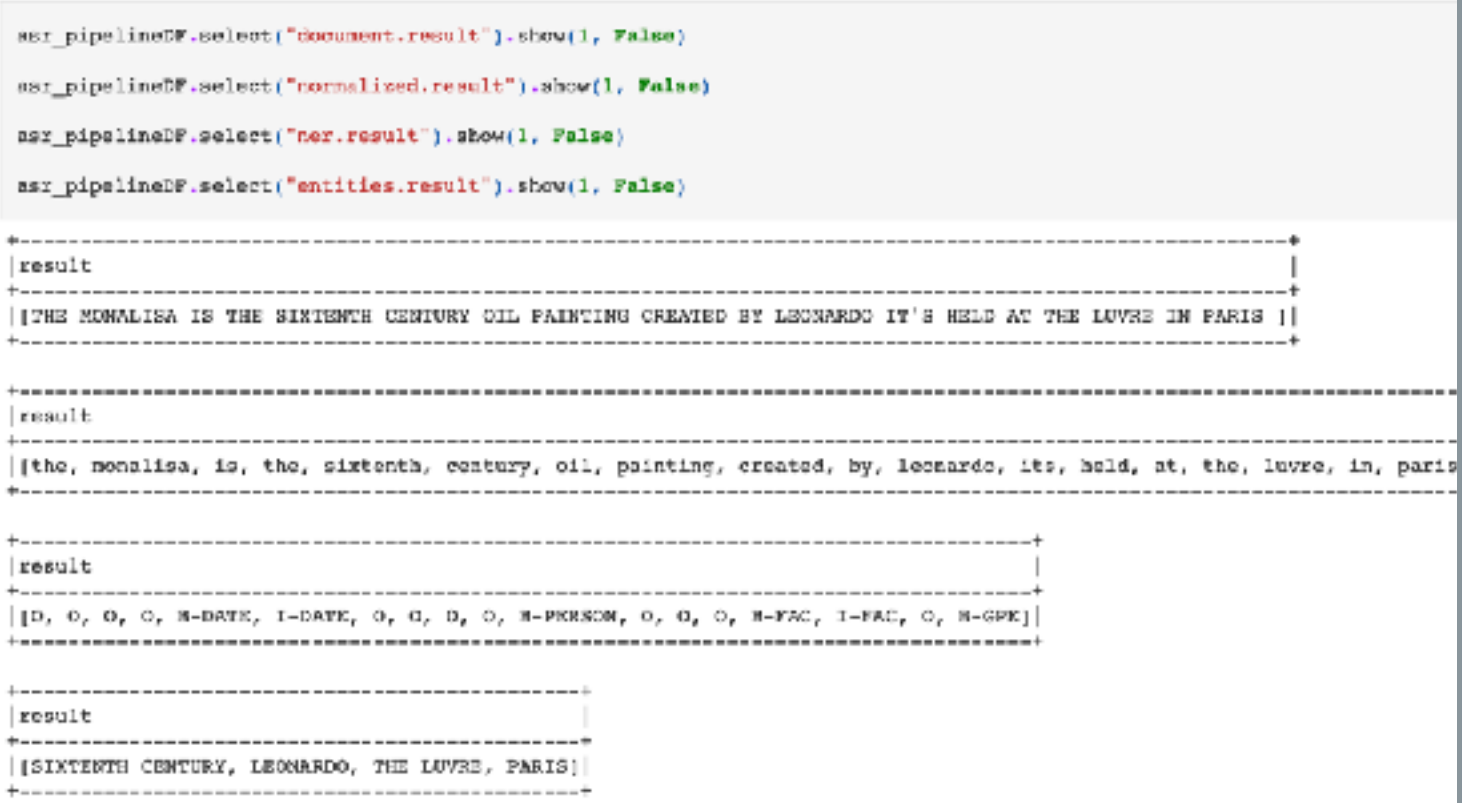

allocations familiales
taxe habitation
aides sociales
classes moyennes
optimisation fiscale
fonction publique
niches fiscales
sécurité sociale
fraude fiscale
élus fonctionnaires
argent public
transition écologique
énergies renouvelables
transition énergétique
autres pays
dépenses publiques
anciens présidents
cour des comptes
assemblée nationale
gilets jaunes
grand débat
services publics
service public
niveau européen
circuits courts


fonction publique
création monétaire
banque privés
pouvoir de création monétaire
banques privées
banque de france
aides sociales
suppression de isf
zones rurales
taux plein
augmentation du nombre
service public
audit citoyen
dette publique
suppression des niches fiscales
niches fiscales suppression
richesse au bénéfice
niches fiscales
paradis fiscaux
sécurité sociale
investissement productif
investissement
coût social
pouvoir achat
évasion fiscale
protection sociale
optimisation fiscale
services publics
optimisation
niveau européen
fraude fiscale
nouvelle disposition
convention judiciaire
transactions financières
produits de luxe
taxe tobin
cotisations sociales
transition écologique
milliards euros
interruption volo
taxation des transactions financières
produits de première nécessité
première nécessité
charges sociales
réchauffement climatique
politique environnementale volontaire
meilleure répartition des richesses
école publique
taux de tva
richesses produites
indexation des retraites
argent public
inflation etat
taux inflation etat
revenu universel
marchés publics


grand plan
votes blancs
allocations fa
premier tour
lassemblée nationale
biens publics
vote blanc
ugement majoritaire
fo
prestations socia
système fiscal
vote obligatoire
transactions financières
administration
tranche supplémentaire
situation régulière
élection présidentielle
mandat présidentiel
prise en compte
élections législatives
conseils municipaux
paradis fiscaux
argent public
proportionnelle intégrale
suffrage universel
évasion fiscale
suffrage universel direct
niches fiscales
tirage au sort
conseil économique
assemblée nationale
union européenne
optimisation fiscale
fraude fiscale
élus locaux
grands groupes
investissement
conseil économique
bonne chose
aide sociale
alimentaires
grandes entreprises
société civile
cour des comptes
produits de première nécessité
classes moyennes
grosses entreprises
conseil constitutionnel
produits de luxe
entreprises françaises
énagères

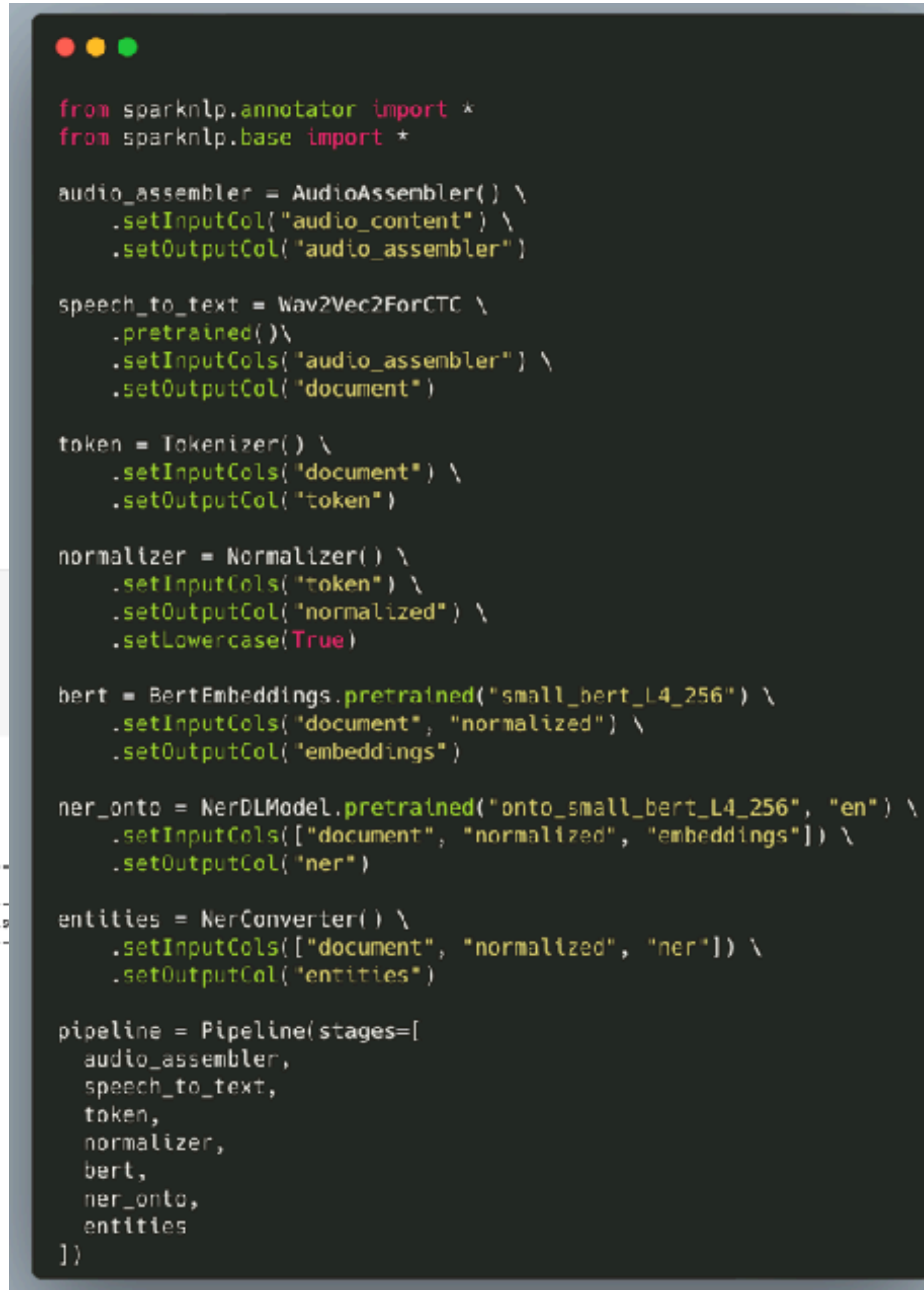
Spark NLP ASR-NER Pipeline

Wav2Vec2, OntoNotes NER, and BERT



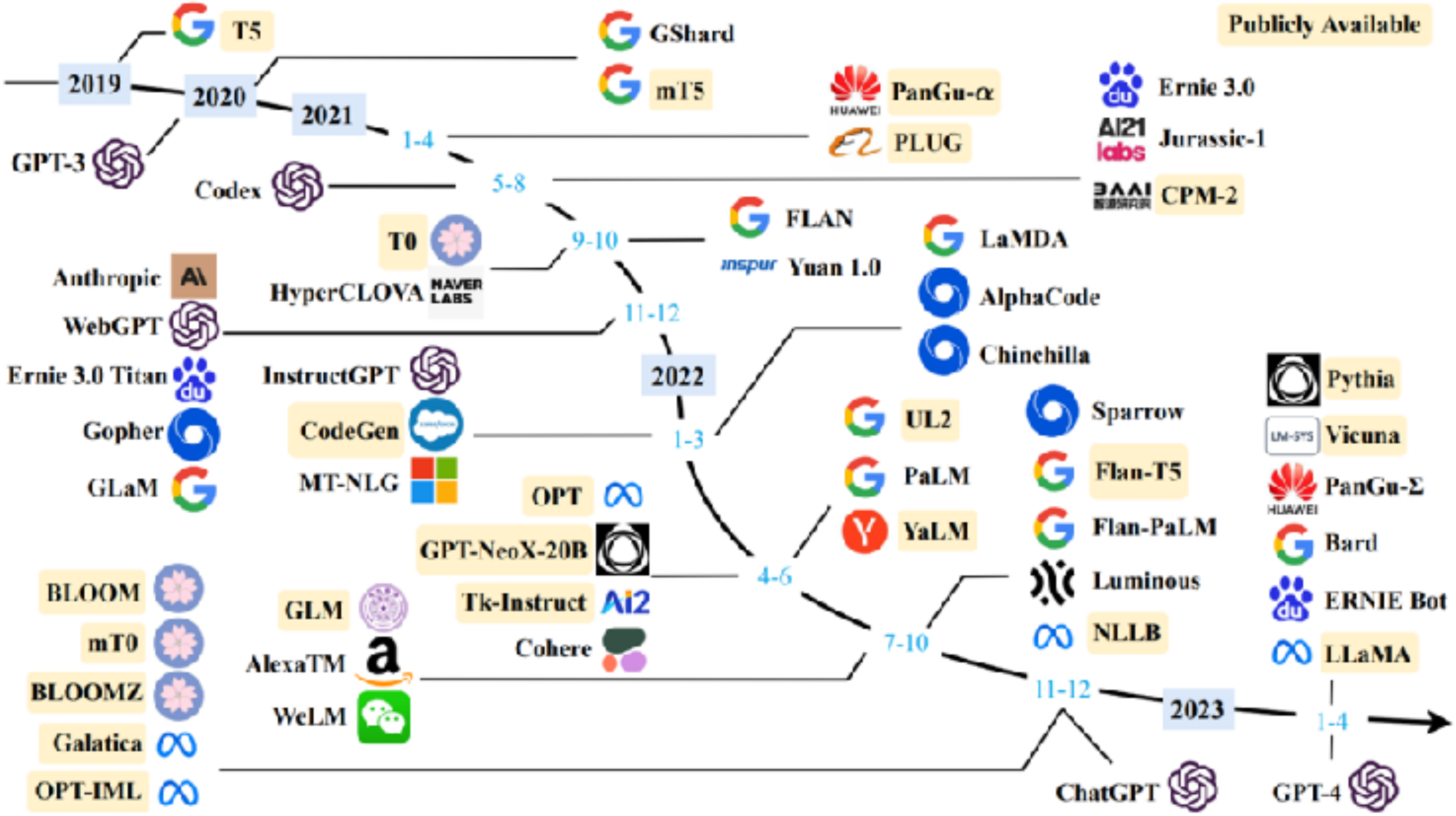
```
asr_pipelineDF.select("document.result").show(1, False)

asr_pipelineDF.select("normalized.result").show(1, False)

asr_pipelineDF.select("ner.result").show(1, False)

asr_pipelineDF.select("entities.result").show(1, False)
```

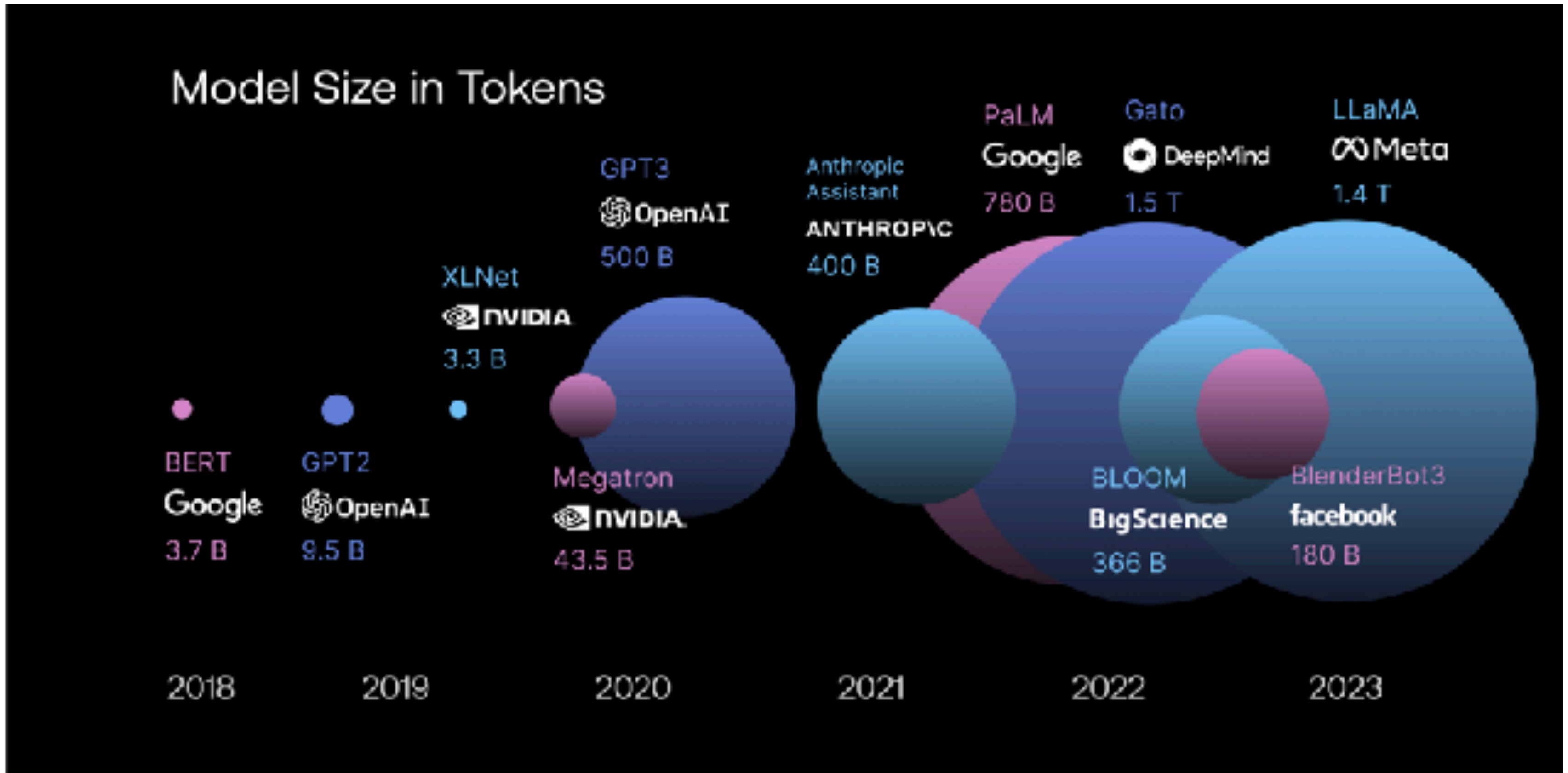
```
+------------------------------------------------------------------------------+
|result                                                                        |
+------------------------------------------------------------------------------+
|[THE MONALISA IS THE SIXTENTH CENTURY OIL PAINTING CREATED BY LEONARDO IT'S HELD AT THE LUVRE IN PARIS ]|
+------------------------------------------------------------------------------+

+------------------------------------------------------------------------------------------
|result
+------------------------------------------------------------------------------------------
|[the, monalisa, is, the, sixtenth, century, oil, painting, created, by, leonardo, its, held, at, the, luvre, in, paris
+------------------------------------------------------------------------------------------

+---------------------------------------------------------------------------+
|result                                                                     |
+---------------------------------------------------------------------------+
|[O, O, O, O, B-DATE, I-DATE, O, O, O, O, B-PERSON, O, O, O, B-FAC, I-FAC, O, B-GPE]|
+---------------------------------------------------------------------------+

+----------------------------------------+
|result                                  |
+----------------------------------------+
|[SIXTENTH CENTURY, LEONARDO, THE LUVRE, PARIS]|
+----------------------------------------+
```



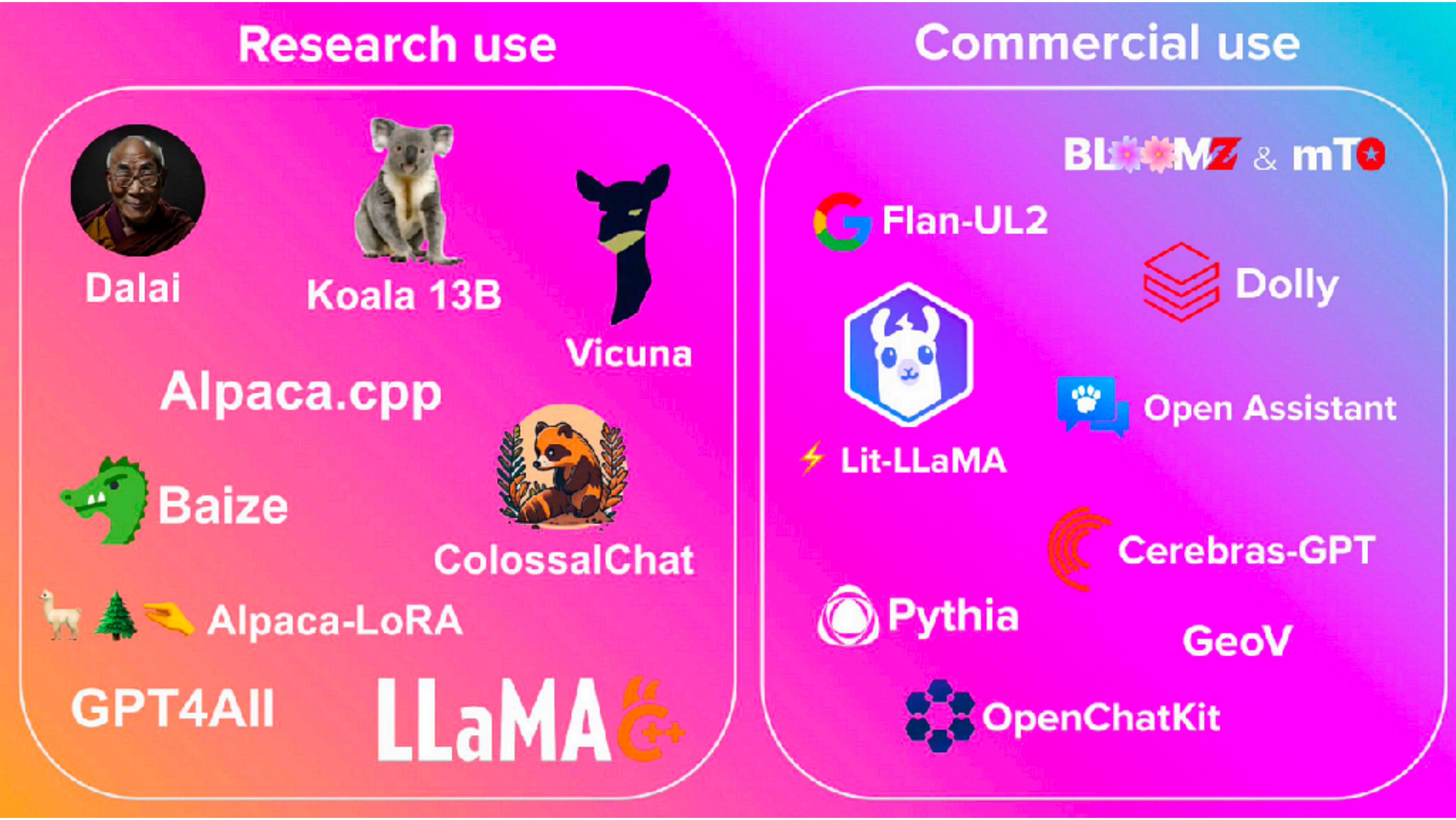
```python
from sparknlp.annotator import *
from sparknlp.base import *

audio_assembler = AudioAssembler() \
    .setInputCol("audio_content") \
    .setOutputCol("audio_assembler")

speech_to_text = Wav2Vec2ForCTC \
    .pretrained()\
    .setInputCols("audio_assembler") \
    .setOutputCol("document")

token = Tokenizer() \
    .setInputCols("document") \
    .setOutputCol("token")

normalizer = Normalizer() \
    .setInputCols("token") \
    .setOutputCol("normalized") \
    .setLowercase(True)

bert = BertEmbeddings.pretrained("small_bert_L4_256") \
    .setInputCols("document", "normalized") \
    .setOutputCol("embeddings")

ner_onto = NerDLModel.pretrained("onto_small_bert_L4_256", "en") \
    .setInputCols(["document", "normalized", "embeddings"]) \
    .setOutputCol("ner")

entities = NerConverter() \
    .setInputCols(["document", "normalized", "ner"]) \
    .setOutputCol("entities")

pipeline = Pipeline(stages=[
  audio_assembler,
  speech_to_text,
  token,
  normalizer,
  bert,
  ner_onto,
  entities
])
```

http://sparknlp.org


Publicly Available
2019
2020
2021
2022
2023
1-4
5-8
9-10
11-12
1-3
4-6
7-10
11-12
1-4
T5
GPT-3
GShard
mT5
Codex
PanGu-α
PLUG
Ernie 3.0
Jurassic-1
CPM-2
T0
HyperCLOVA
FLAN
Yuan 1.0
Anthropic
WebGPT
Ernie 3.0 Titan
Gopher
GLaM
InstructGPT
CodeGen
MT-NLG
LaMDA
AlphaCode
Chinchilla
UL2
PaLM
YaLM
Sparrow
Flan-T5
Flan-PaLM
Luminous
NLLB
Pythia
Vicuna
PanGu-Σ
Bard
ERNIE Bot
LLaMA
OPT
GPT-NeoX-20B
Tk-Instruct
Cohere
GLM
AlexaTM
WeLM
BLOOM
mT0
BLOOMZ
Galatica
OPT-IML
ChatGPT
GPT-4

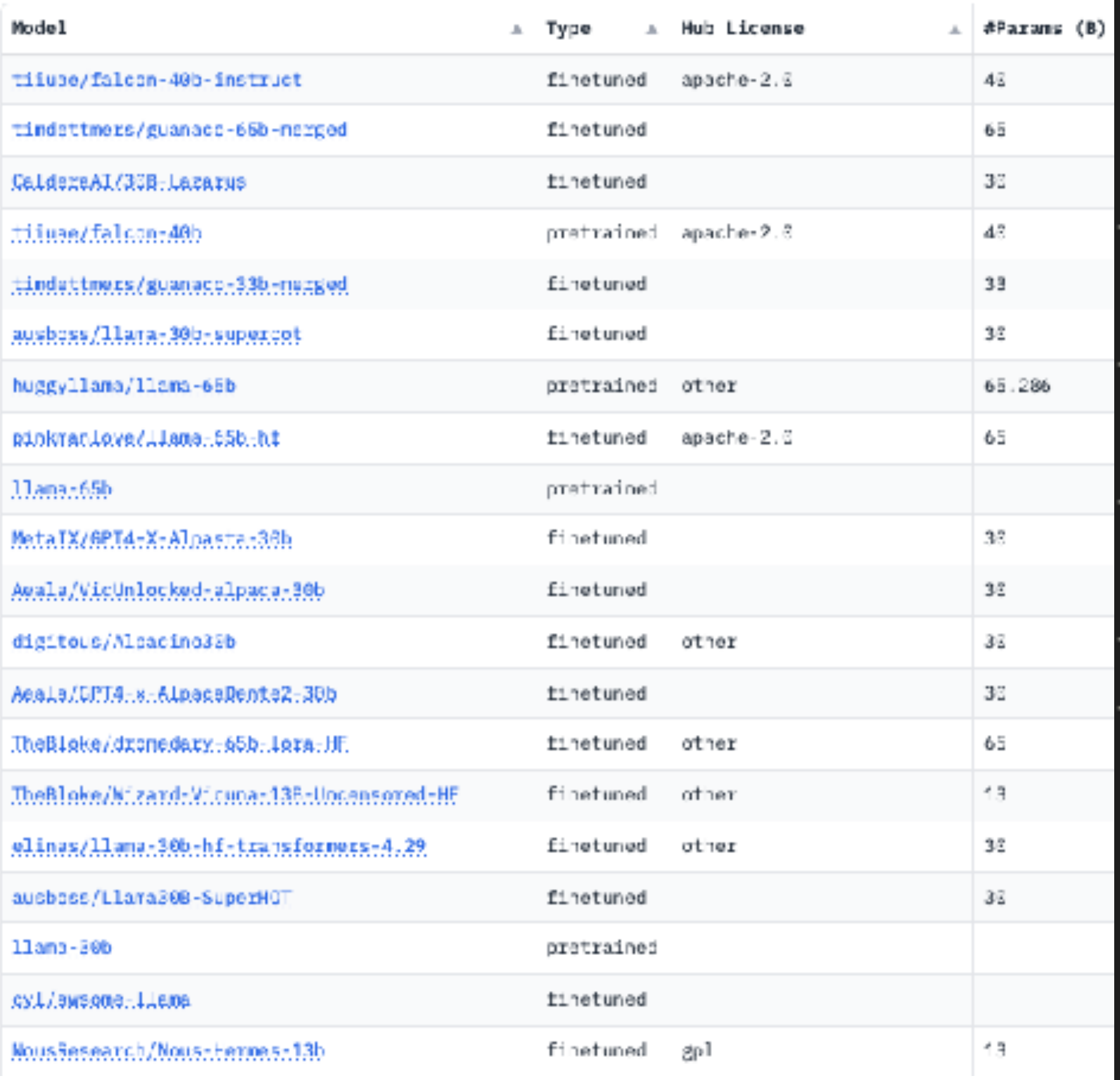

Model Size in Tokens
BERT
Google
3.7 B
GPT2
OpenAI
9.5 B
XLNet
NVIDIA
3.3 B
GPT3
OpenAI
500 B
Megatron
NVIDIA
43.5 B
Anthropic Assistant
ANTHROPIC
400 B
PaLM
Google
780 B
Gato
DeepMind
1.5 T
BLOOM
BigScience
366 B
LLaMA
Meta
1.4 T
BlenderBot3
facebook
180 B
2018
2019
2020
2021
2022
2023

Large Language Models



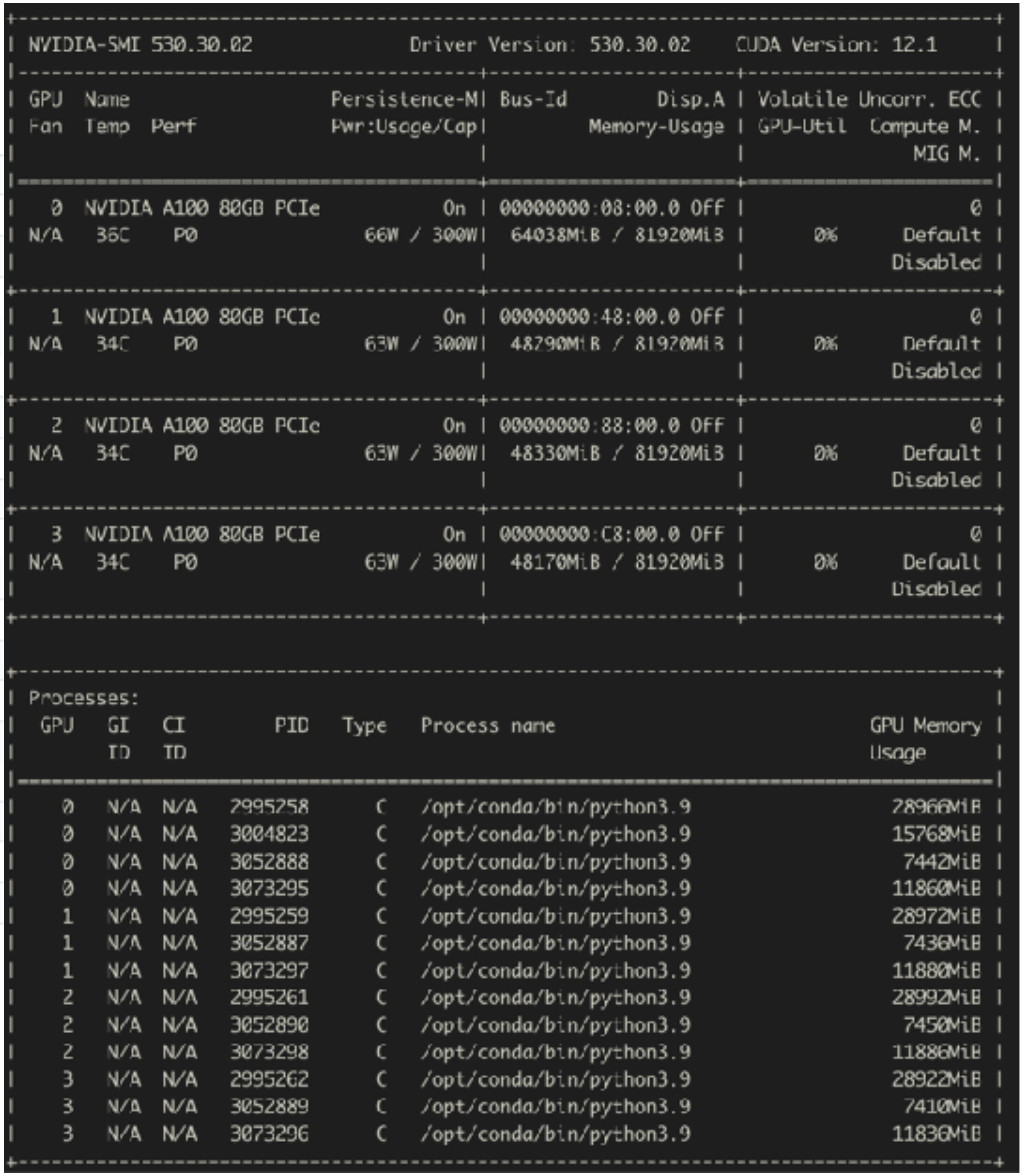
cnrs Institut des Systèmes Complexes Paris Île-de-France

LLM Leaderboard

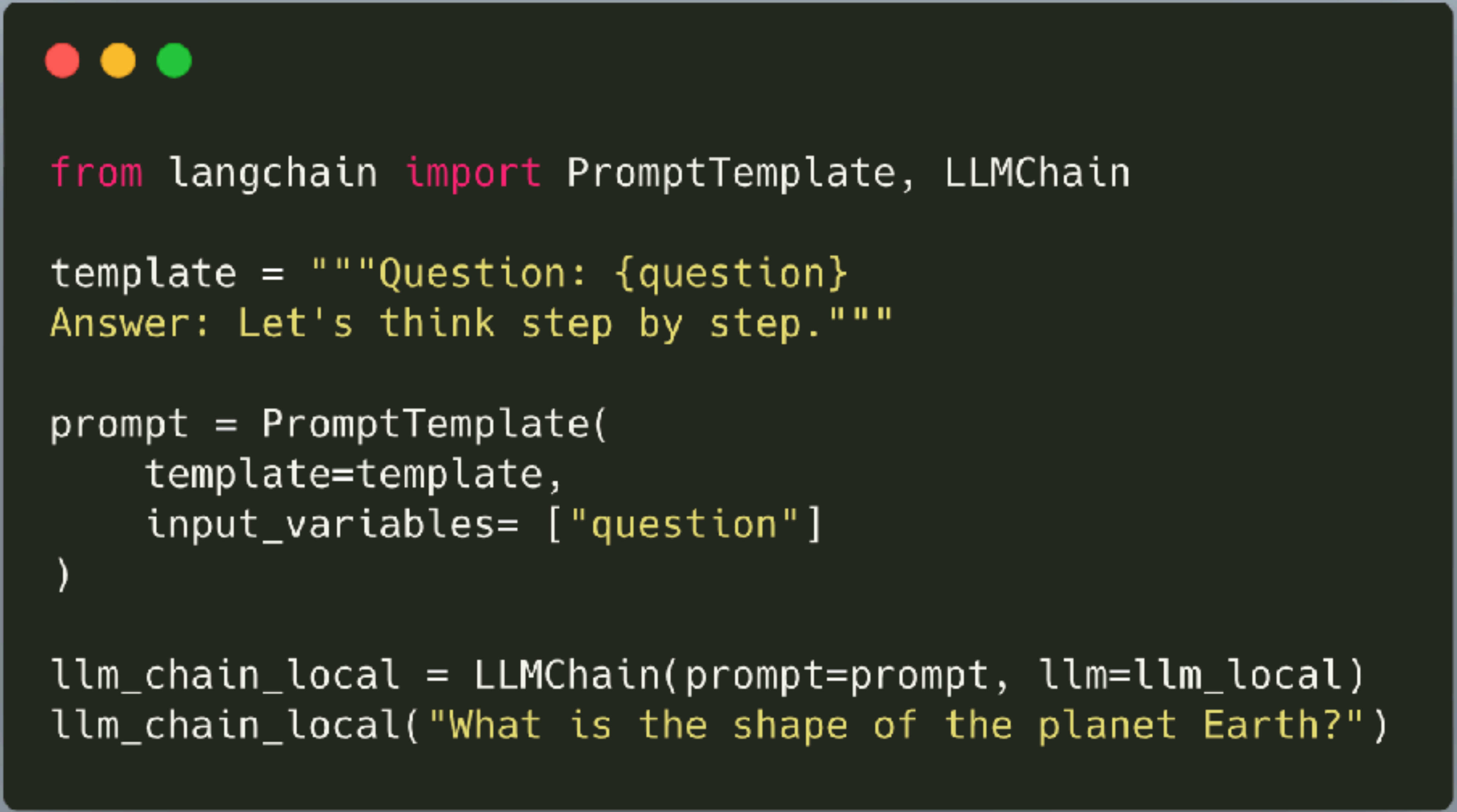


Model	Type	Hub License	#Params (B)
tiiuae/falcon-40b-instruct	finetuned	apache-2.0	42
timdettmers/guanaco-65b-merged	finetuned		65
CalderaAI/30B-Lazarus	finetuned		32
tiiuae/falcon-40b	pretrained	apache-2.0	42
timdettmers/guanaco-33b-merged	finetuned		33
ausboss/llama-30b-supercot	finetuned		32
huggyllama/llama-65b	pretrained	other	65.286
pinkmanlove/llama-65b-hf	finetuned	apache-2.0	65
llama-65b	pretrained		
MetaIX/GPT4-X-Alpasta-30b	finetuned		33
Aeala/VicUnlocked-alpaca-30b	finetuned		32
digitous/Alpacino30b	finetuned	other	32
Aeala/GPT4-x-AlpacaDente2-30b	finetuned		32
TheBloke/dromedary-65b-lora-HF	finetuned	other	65
TheBloke/Wizard-Vicuna-13B-Uncensored-HF	finetuned	other	13
elinas/llama-30b-hf-transformers-4.29	finetuned	other	32
ausboss/Llama30B-SuperHOT	finetuned		32
llama-30b	pretrained		
cyl/awsome-llama	finetuned		
NousResearch/Nous-Hermes-13b	finetuned	gpl	13

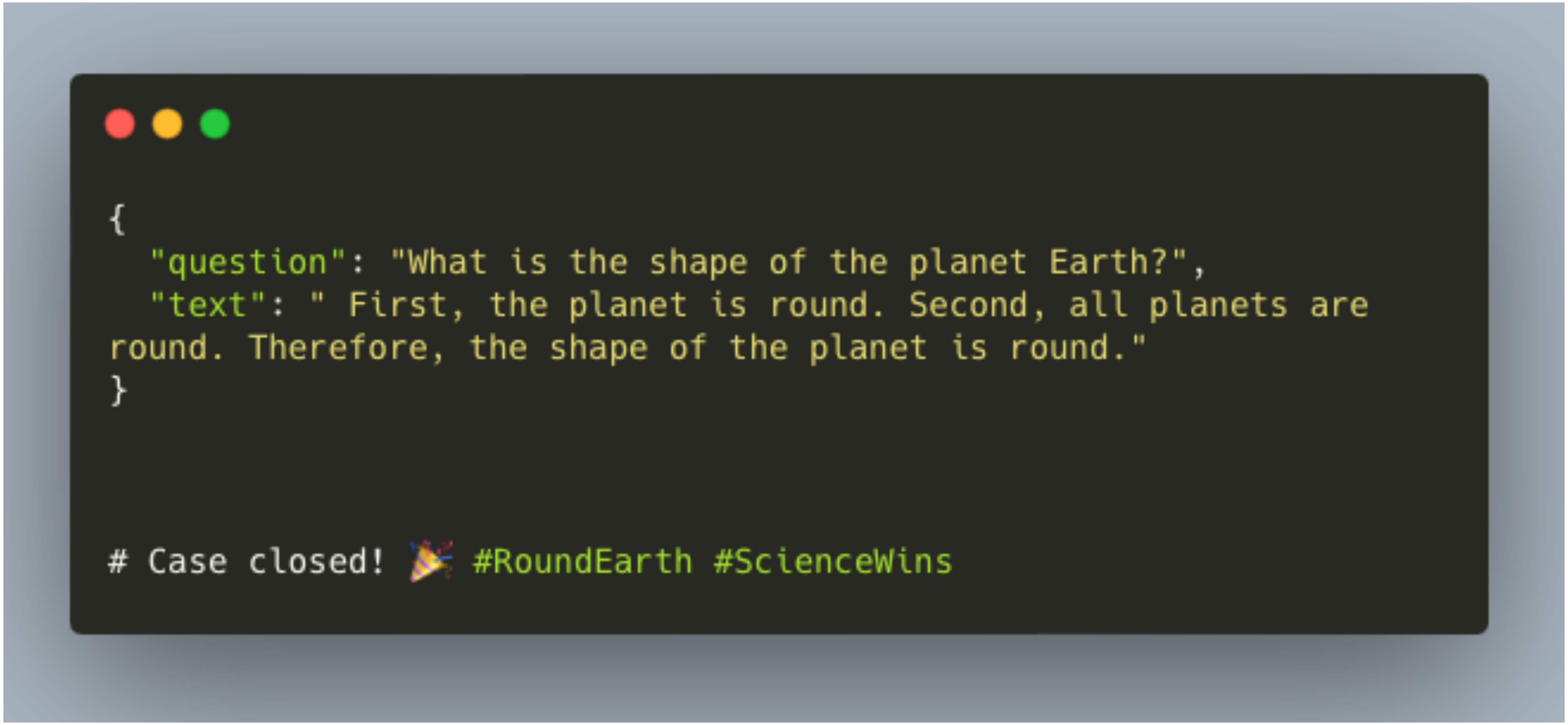
https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard



```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 530.30.02            Driver Version: 530.30.02    CUDA Version: 12.1     |
|-----------------------------------------+----------------------+----------------------+
| GPU  Name                  Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf            Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|                                         |                      |               MIG M. |
|=========================================+======================+======================|
|   0  NVIDIA A100 80GB PCIe           On | 00000000:08:00.0 Off |                    0 |
| N/A   36C    P0             66W / 300W|  64038MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   1  NVIDIA A100 80GB PCIe           On | 00000000:48:00.0 Off |                    0 |
| N/A   34C    P0             63W / 300W|  48290MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   2  NVIDIA A100 80GB PCIe           On | 00000000:88:00.0 Off |                    0 |
| N/A   34C    P0             63W / 300W|  48330MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+
|   3  NVIDIA A100 80GB PCIe           On | 00000000:C8:00.0 Off |                    0 |
| N/A   34C    P0             63W / 300W|  48170MiB / 81920MiB |      0%      Default |
|                                         |                      |             Disabled |
+-----------------------------------------+----------------------+----------------------+

+---------------------------------------------------------------------------------------+
| Processes:                                                                            |
|  GPU   GI   CI        PID   Type   Process name                            GPU Memory |
|        ID   ID                                                             Usage      |
|=======================================================================================|
|    0   N/A  N/A    2995258      C   /opt/conda/bin/python3.9                  28966MiB |
|    0   N/A  N/A    3004823      C   /opt/conda/bin/python3.9                  15768MiB |
|    0   N/A  N/A    3052888      C   /opt/conda/bin/python3.9                   7442MiB |
|    0   N/A  N/A    3073295      C   /opt/conda/bin/python3.9                  11860MiB |
|    1   N/A  N/A    2995259      C   /opt/conda/bin/python3.9                  28972MiB |
|    1   N/A  N/A    3052887      C   /opt/conda/bin/python3.9                   7436MiB |
|    1   N/A  N/A    3073297      C   /opt/conda/bin/python3.9                  11880MiB |
|    2   N/A  N/A    2995261      C   /opt/conda/bin/python3.9                  28992MiB |
|    2   N/A  N/A    3052890      C   /opt/conda/bin/python3.9                   7450MiB |
|    2   N/A  N/A    3073298      C   /opt/conda/bin/python3.9                  11886MiB |
|    3   N/A  N/A    2995262      C   /opt/conda/bin/python3.9                  28922MiB |
|    3   N/A  N/A    3052889      C   /opt/conda/bin/python3.9                   7410MiB |
|    3   N/A  N/A    3073296      C   /opt/conda/bin/python3.9                  11836MiB |
+---------------------------------------------------------------------------------------+
```



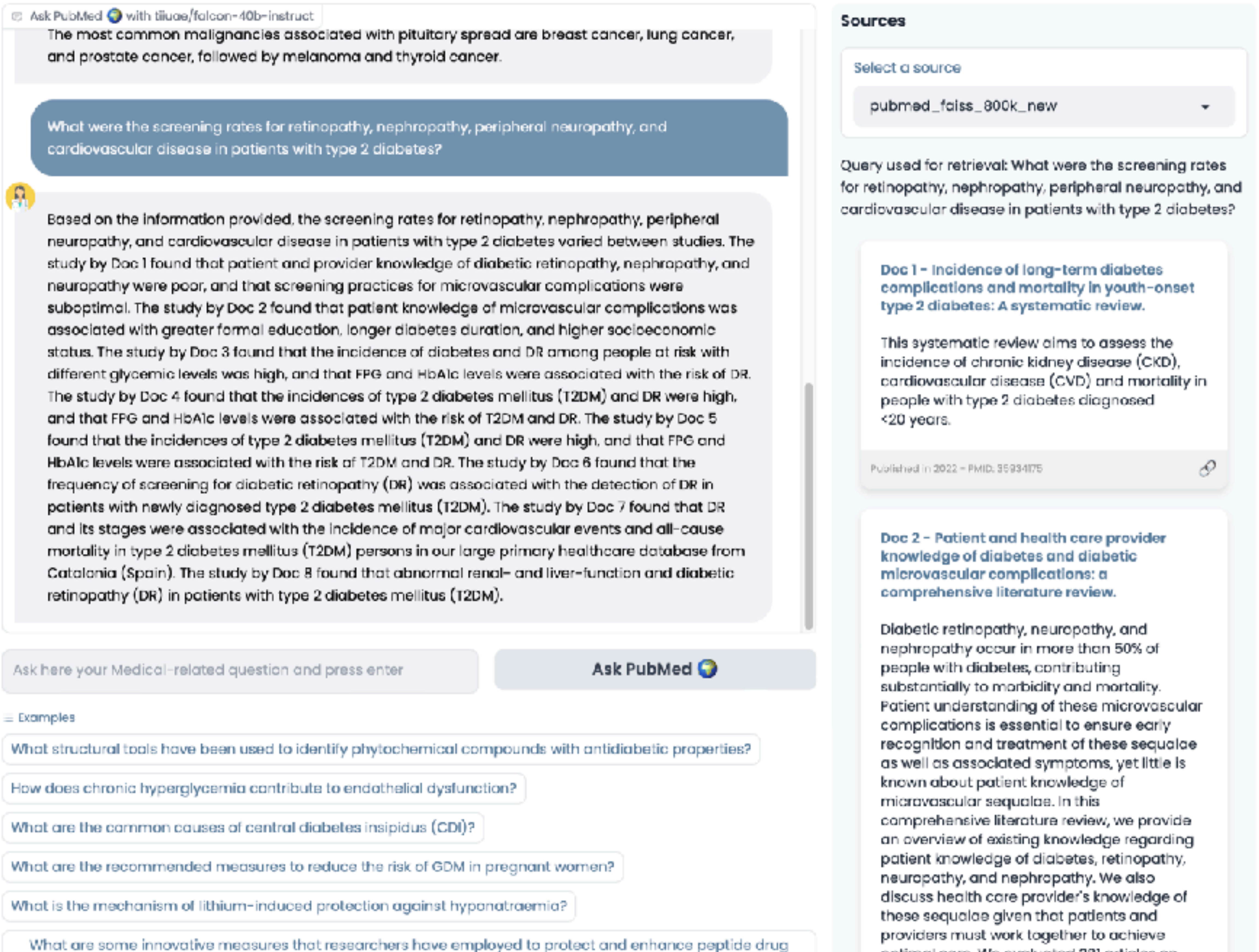
```python
from langchain import PromptTemplate, LLMChain

template = """Question: {question}
Answer: Let's think step by step."""

prompt = PromptTemplate(
    template=template,
    input_variables= ["question"]
)

llm_chain_local = LLMChain(prompt=prompt, llm=llm_local)
llm_chain_local("What is the shape of the planet Earth?")
```



```
{
  "question": "What is the shape of the planet Earth?",
  "text": " First, the planet is round. Second, all planets are
round. Therefore, the shape of the planet is round."
}



# Case closed! 🎉 #RoundEarth #ScienceWins
```

Ask **PubMed**

falcon-40b-instruct



The most common malignancies associated with pituitary spread are breast cancer, lung cancer, and prostate cancer, followed by melanoma and thyroid cancer.

What were the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes?

Based on the information provided, the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes varied between studies. The study by Doc 1 found that patient and provider knowledge of diabetic retinopathy, nephropathy, and neuropathy were poor, and that screening practices for microvascular complications were suboptimal. The study by Doc 2 found that patient knowledge of microvascular complications was associated with greater formal education, longer diabetes duration, and higher socioeconomic status. The study by Doc 3 found that the incidence of diabetes and DR among people at risk with different glycemic levels was high, and that FPG and HbA1c levels were associated with the risk of DR. The study by Doc 4 found that the incidences of type 2 diabetes mellitus (T2DM) and DR were high, and that FPG and HbA1c levels were associated with the risk of T2DM and DR. The study by Doc 5 found that the incidences of type 2 diabetes mellitus (T2DM) and DR were high, and that FPG and HbA1c levels were associated with the risk of T2DM and DR. The study by Doc 6 found that the frequency of screening for diabetic retinopathy (DR) was associated with the detection of DR in patients with newly diagnosed type 2 diabetes mellitus (T2DM). The study by Doc 7 found that DR and its stages were associated with the incidence of major cardiovascular events and all-cause mortality in type 2 diabetes mellitus (T2DM) persons in our large primary healthcare database from Catalonia (Spain). The study by Doc 8 found that abnormal renal- and liver-function and diabetic retinopathy (DR) in patients with type 2 diabetes mellitus (T2DM).

What structural tools have been used to identify phytochemical compounds with antidiabetic properties?

How does chronic hyperglycemia contribute to endothelial dysfunction?

What are the common causes of central diabetes insipidus (CDI)?

What are the recommended measures to reduce the risk of GDM in pregnant women?

What is the mechanism of lithium-induced protection against hyponatraemia?

What are some innovative measures that researchers have employed to protect and enhance peptide drug

Sources

Select a source: pubmed_faiss_800k_new

Query used for retrieval: What were the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes?

Doc 1 - Incidence of long-term diabetes complications and mortality in youth-onset type 2 diabetes: A systematic review.

This systematic review aims to assess the incidence of chronic kidney disease (CKD), cardiovascular disease (CVD) and mortality in people with type 2 diabetes diagnosed <20 years.

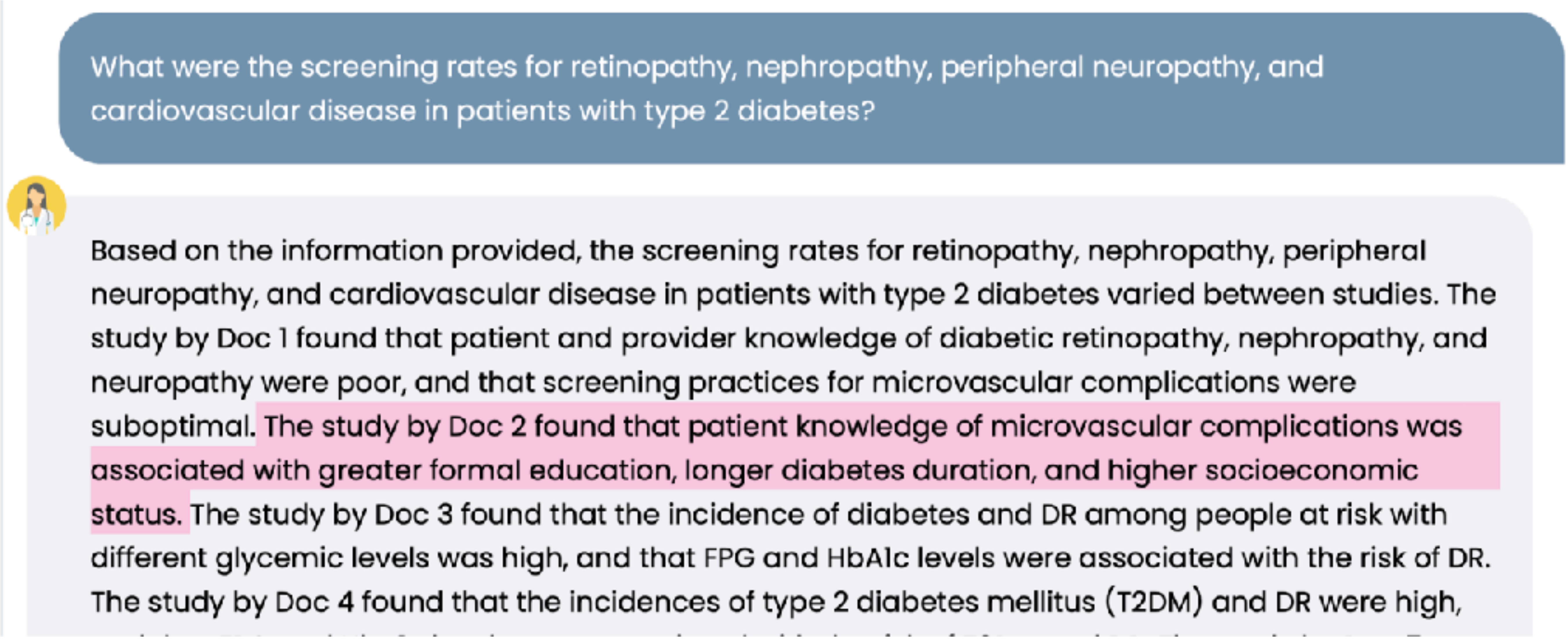
Published in 2022 - PMID: 35934175

Doc 2 - Patient and health care provider knowledge of diabetes and diabetic microvascular complications: a comprehensive literature review.

Diabetic retinopathy, neuropathy, and nephropathy occur in more than 50% of people with diabetes, contributing substantially to morbidity and mortality. Patient understanding of these microvascular complications is essential to ensure early recognition and treatment of these sequelae as well as associated symptoms, yet little is known about patient knowledge of microvascular sequelae. In this comprehensive literature review, we provide an overview of existing knowledge regarding patient knowledge of diabetes, retinopathy, neuropathy, and nephropathy. We also discuss health care provider's knowledge of these sequelae given that patients and providers must work together to achieve

Ask **PubMed**

falcon-40b-instruct

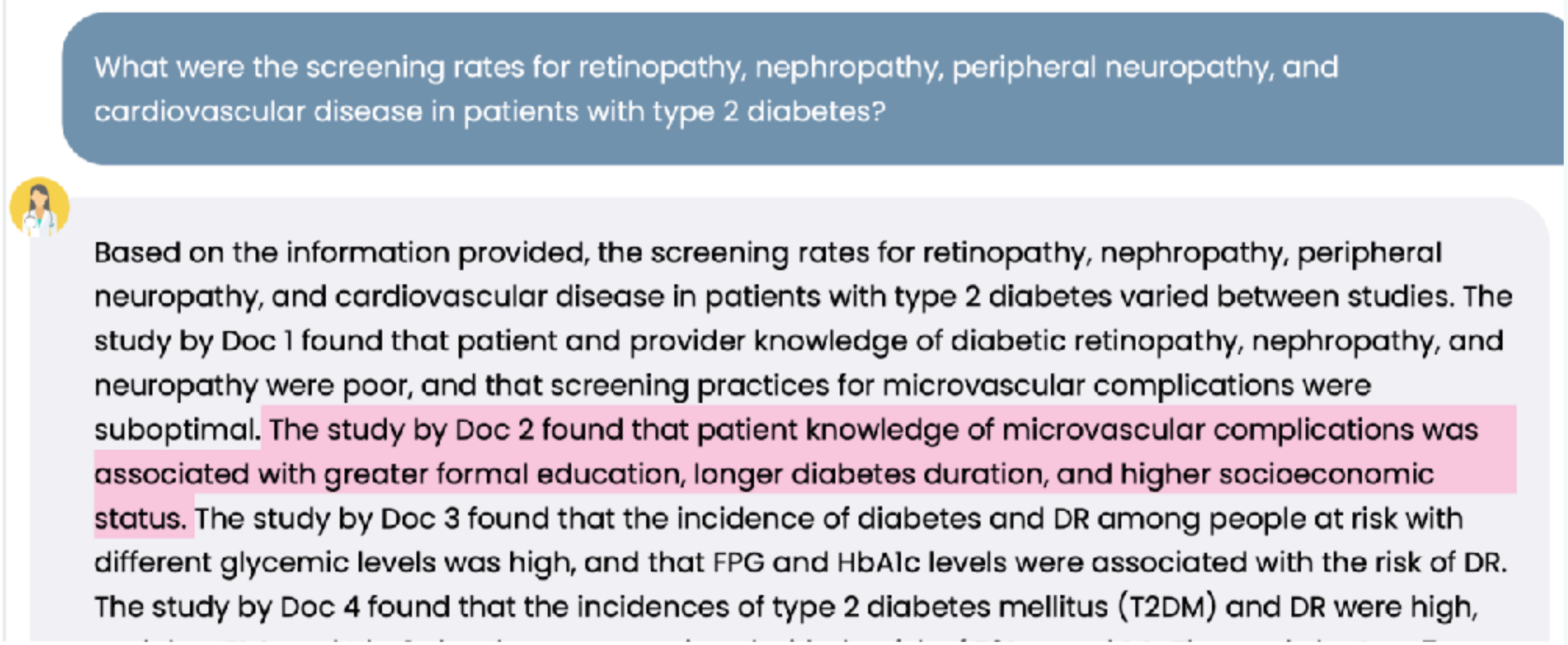


What were the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes?

Based on the information provided, the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes varied between studies. The study by Doc 1 found that patient and provider knowledge of diabetic retinopathy, nephropathy, and neuropathy were poor, and that screening practices for microvascular complications were suboptimal. The study by Doc 2 found that patient knowledge of microvascular complications was associated with greater formal education, longer diabetes duration, and higher socioeconomic status. The study by Doc 3 found that the incidence of diabetes and DR among people at risk with different glycemic levels was high, and that FPG and HbA1c levels were associated with the risk of DR. The study by Doc 4 found that the incidences of type 2 diabetes mellitus (T2DM) and DR were high,

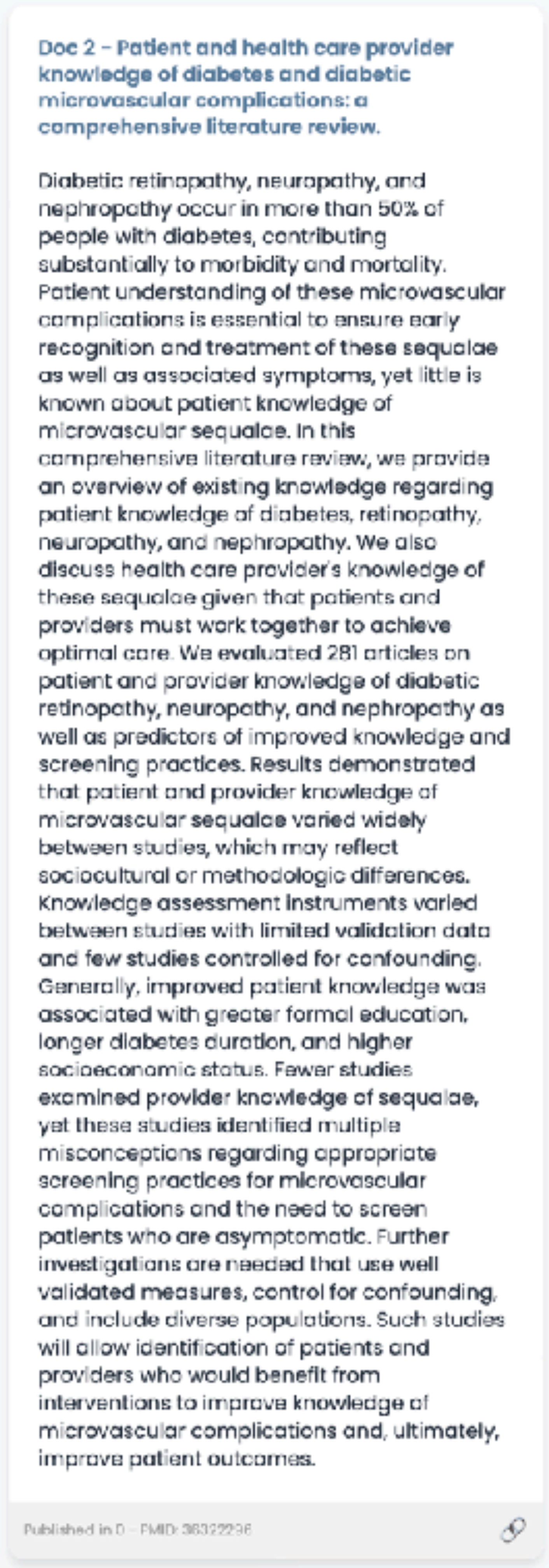
Ask **PubMed**

falcon-40b-instruct



What were the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes?

Based on the information provided, the screening rates for retinopathy, nephropathy, peripheral neuropathy, and cardiovascular disease in patients with type 2 diabetes varied between studies. The study by Doc 1 found that patient and provider knowledge of diabetic retinopathy, nephropathy, and neuropathy were poor, and that screening practices for microvascular complications were suboptimal. The study by Doc 2 found that patient knowledge of microvascular complications was associated with greater formal education, longer diabetes duration, and higher socioeconomic status. The study by Doc 3 found that the incidence of diabetes and DR among people at risk with different glycemic levels was high, and that FPG and HbA1c levels were associated with the risk of DR. The study by Doc 4 found that the incidences of type 2 diabetes mellitus (T2DM) and DR were high,

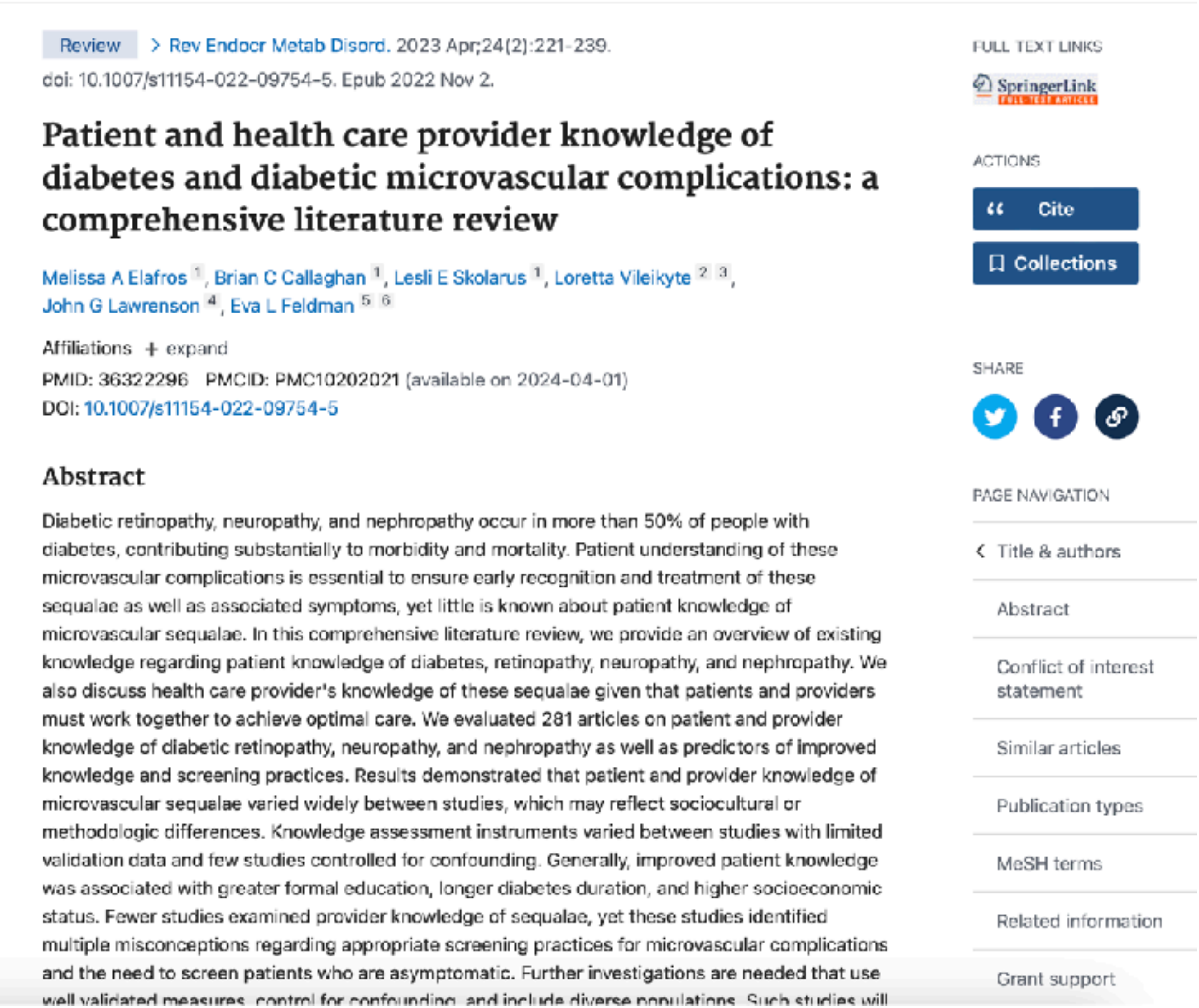


Doc 2 – Patient and health care provider knowledge of diabetes and diabetic microvascular complications: a comprehensive literature review.

Diabetic retinopathy, neuropathy, and nephropathy occur in more than 50% of people with diabetes, contributing substantially to morbidity and mortality. Patient understanding of these microvascular complications is essential to ensure early recognition and treatment of these sequalae as well as associated symptoms, yet little is known about patient knowledge of microvascular sequalae. In this comprehensive literature review, we provide an overview of existing knowledge regarding patient knowledge of diabetes, retinopathy, neuropathy, and nephropathy. We also discuss health care provider's knowledge of these sequalae given that patients and providers must work together to achieve optimal care. We evaluated 281 articles on patient and provider knowledge of diabetic retinopathy, neuropathy, and nephropathy as well as predictors of improved knowledge and screening practices. Results demonstrated that patient and provider knowledge of microvascular sequalae varied widely between studies, which may reflect sociocultural or methodologic differences. Knowledge assessment instruments varied between studies with limited validation data and few studies controlled for confounding. Generally, improved patient knowledge was associated with greater formal education, longer diabetes duration, and higher socioeconomic status. Fewer studies examined provider knowledge of sequalae, yet these studies identified multiple misconceptions regarding appropriate screening practices for microvascular complications and the need to screen patients who are asymptomatic. Further investigations are needed that use well validated measures, control for confounding, and include diverse populations. Such studies will allow identification of patients and providers who would benefit from interventions to improve knowledge of microvascular complications and, ultimately, improve patient outcomes.

Published in D – PMID: 36322236

cnrs Institut des Systèmes Complexes Paris Île-de-France



Review > Rev Endocr Metab Disord. 2023 Apr;24(2):221-239.
doi: 10.1007/s11154-022-09754-5. Epub 2022 Nov 2.

Patient and health care provider knowledge of diabetes and diabetic microvascular complications: a comprehensive literature review

Melissa A Elafros [1], Brian C Callaghan [1], Lesli E Skolarus [1], Loretta Vileikyte [2] [3], John G Lawrenson [4], Eva L Feldman [5] [6]

Affiliations + expand

PMID: 36322296 PMCID: PMC10202021 (available on 2024-04-01)
DOI: 10.1007/s11154-022-09754-5

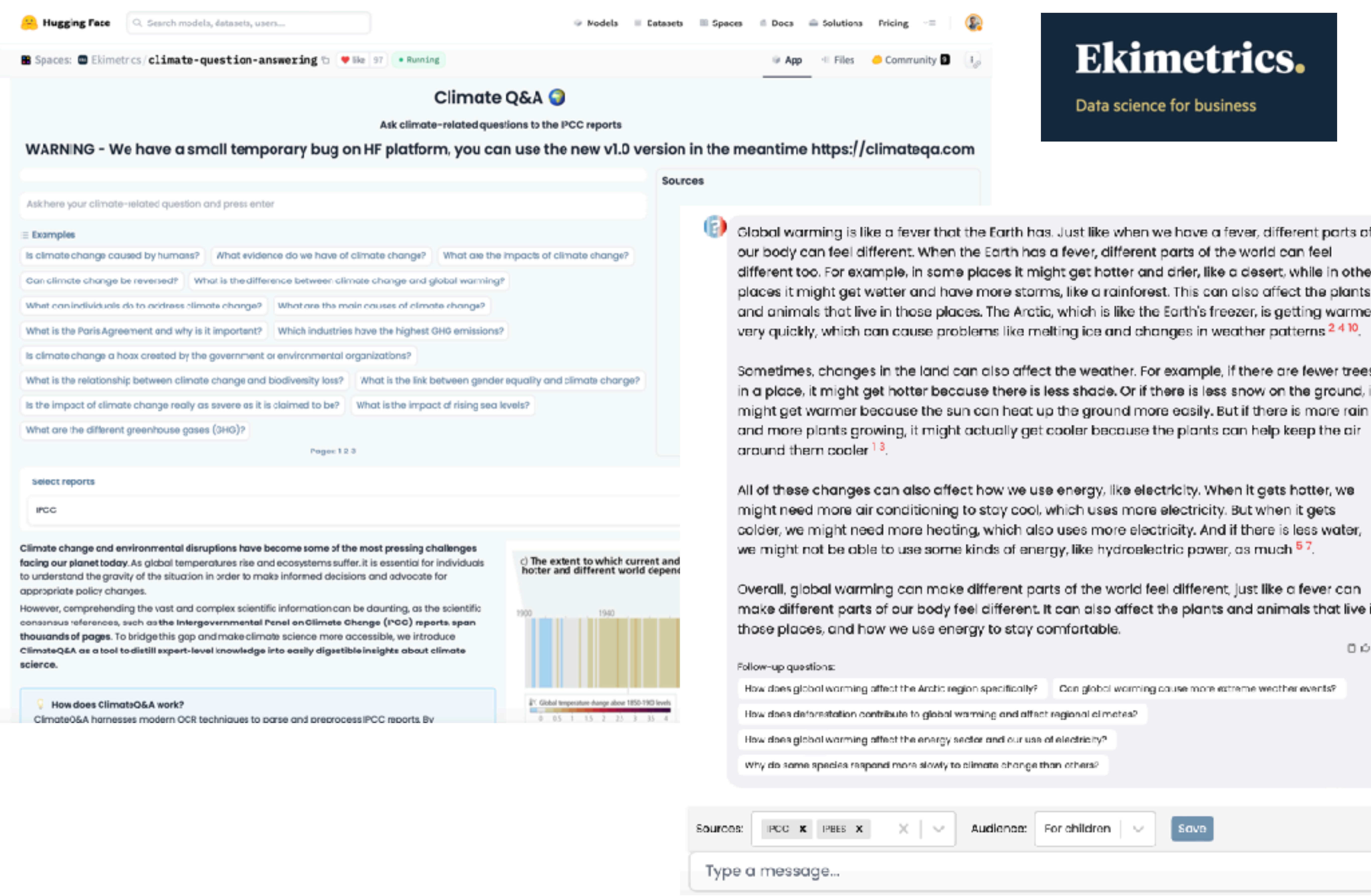
Abstract

Diabetic retinopathy, neuropathy, and nephropathy occur in more than 50% of people with diabetes, contributing substantially to morbidity and mortality. Patient understanding of these microvascular complications is essential to ensure early recognition and treatment of these sequalae as well as associated symptoms, yet little is known about patient knowledge of microvascular sequalae. In this comprehensive literature review, we provide an overview of existing knowledge regarding patient knowledge of diabetes, retinopathy, neuropathy, and nephropathy. We also discuss health care provider's knowledge of these sequalae given that patients and providers must work together to achieve optimal care. We evaluated 281 articles on patient and provider knowledge of diabetic retinopathy, neuropathy, and nephropathy as well as predictors of improved knowledge and screening practices. Results demonstrated that patient and provider knowledge of microvascular sequalae varied widely between studies, which may reflect sociocultural or methodologic differences. Knowledge assessment instruments varied between studies with limited validation data and few studies controlled for confounding. Generally, improved patient knowledge was associated with greater formal education, longer diabetes duration, and higher socioeconomic status. Fewer studies examined provider knowledge of sequalae, yet these studies identified multiple misconceptions regarding appropriate screening practices for microvascular complications and the need to screen patients who are asymptomatic. Further investigations are needed that use well validated measures, control for confounding, and include diverse populations. Such studies will

cnrs Institut des Systèmes Complexes

Ekimetrics.

Data science for business



Climate Q&A 🌍

Ask climate-related questions to the IPCC reports

WARNING - We have a small temporary bug on HF platform, you can use the new v1.0 version in the meantime https://climateqa.com

Ask here your climate-related question and press enter

Examples

- Is climate change caused by humans?
- What evidence do we have of climate change?
- What are the impacts of climate change?
- Can climate change be reversed?
- What is the difference between climate change and global warming?
- What can individuals do to address climate change?
- What are the main causes of climate change?
- What is the Paris Agreement and why is it important?
- Which industries have the highest GHG emissions?
- Is climate change a hoax created by the government or environmental organizations?
- What is the relationship between climate change and biodiversity loss?
- What is the link between gender equality and climate change?
- Is the impact of climate change really as severe as it is claimed to be?
- What is the impact of rising sea levels?
- What are the different greenhouse gases (GHG)?

Pages: 1 2 3

Select reports: IPCC

Climate change and environmental disruptions have become some of the most pressing challenges facing our planet today. As global temperatures rise and ecosystems suffer, it is essential for individuals to understand the gravity of the situation in order to make informed decisions and advocate for appropriate policy changes.

However, comprehending the vast and complex scientific information can be daunting, as the scientific consensus references, such as **the Intergovernmental Panel on Climate Change (IPCC) reports, span thousands of pages**. To bridge this gap and make climate science more accessible, we introduce **ClimateQ&A as a tool to distill expert-level knowledge into easily digestible insights about climate science.**

How does ClimateQ&A work?

ClimateQ&A harnesses modern OCR techniques to parse and preprocess IPCC reports. By

c) The extent to which current and
hotter and different world depend
1900
1940
Global temperature change above 1850-1900 levels
0
0.5
1
1.5
2
2.5
3
3.5
4

Sources

Global warming is like a fever that the Earth has. Just like when we have a fever, different parts o our body can feel different. When the Earth has a fever, different parts of the world can feel different too. For example, in some places it might get hotter and drier, like a desert, while in othe places it might get wetter and have more storms, like a rainforest. This can also affect the plants and animals that live in those places. The Arctic, which is like the Earth's freezer, is getting warme very quickly, which can cause problems like melting ice and changes in weather patterns[2 4 10].

Sometimes, changes in the land can also affect the weather. For example, if there are fewer tree in a place, it might get hotter because there is less shade. Or if there is less snow on the ground, i might get warmer because the sun can heat up the ground more easily. But if there is more rain and more plants growing, it might actually get cooler because the plants can help keep the air around them cooler[1 3].

All of these changes can also affect how we use energy, like electricity. When it gets hotter, we might need more air conditioning to stay cool, which uses more electricity. But when it gets colder, we might need more heating, which also uses more electricity. And if there is less water, we might not be able to use some kinds of energy, like hydroelectric power, as much[5 7].

Overall, global warming can make different parts of the world feel different, just like a fever can make different parts of our body feel different. It can also affect the plants and animals that live i those places, and how we use energy to stay comfortable.

Follow-up questions:

- How does global warming affect the Arctic region specifically?
- Can global warming cause more extreme weather events?
- How does deforestation contribute to global warming and affect regional climates?
- How does global warming affect the energy sector and our use of electricity?
- Why do some species respond more slowly to climate change than others?

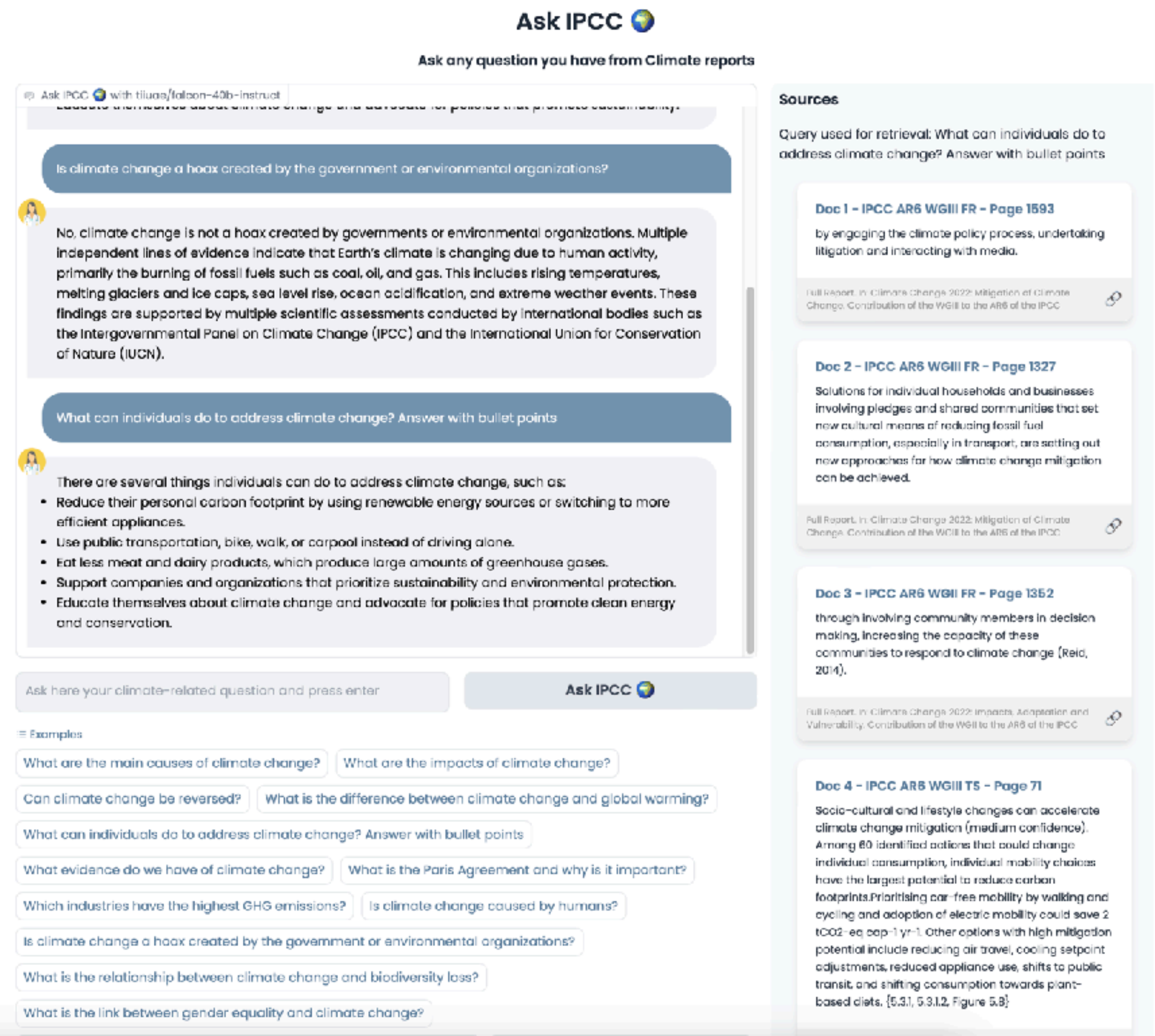
Sources: IPCC, IPBES

Audience: For children

Institut des Systèmes Complexes

Paris Île-de-France

cnrs



Ask IPCC 🌍

Ask any question you have from Climate reports

Educate themselves about climate change and advocate for policies that promote sustainability.

Is climate change a hoax created by the government or environmental organizations?

No, climate change is not a hoax created by governments or environmental organizations. Multiple independent lines of evidence indicate that Earth's climate is changing due to human activity, primarily the burning of fossil fuels such as coal, oil, and gas. This includes rising temperatures, melting glaciers and ice caps, sea level rise, ocean acidification, and extreme weather events. These findings are supported by multiple scientific assessments conducted by international bodies such as the Intergovernmental Panel on Climate Change (IPCC) and the International Union for Conservation of Nature (IUCN).

What can individuals do to address climate change? Answer with bullet points

There are several things individuals can do to address climate change, such as:

- Reduce their personal carbon footprint by using renewable energy sources or switching to more efficient appliances.
- Use public transportation, bike, walk, or carpool instead of driving alone.
- Eat less meat and dairy products, which produce large amounts of greenhouse gases.
- Support companies and organizations that prioritize sustainability and environmental protection.
- Educate themselves about climate change and advocate for policies that promote clean energy and conservation.

What are the main causes of climate change?
What are the impacts of climate change?
Can climate change be reversed?
What is the difference between climate change and global warming?
What can individuals do to address climate change? Answer with bullet points
What evidence do we have of climate change?
What is the Paris Agreement and why is it important?
Which industries have the highest GHG emissions?
Is climate change caused by humans?
Is climate change a hoax created by the government or environmental organizations?
What is the relationship between climate change and biodiversity loss?
What is the link between gender equality and climate change?

Sources

Query used for retrieval: What can individuals do to address climate change? Answer with bullet points

Doc 1 - IPCC AR6 WGIII FR - Page 1593

by engaging the climate policy process, undertaking litigation and interacting with media.

Full Report. In: Climate Change 2022: Mitigation of Climate Change. Contribution of the WGIII to the AR6 of the IPCC

Doc 2 - IPCC AR6 WGIII FR - Page 1327

Solutions for individual households and businesses involving pledges and shared communities that set new cultural means of reducing fossil fuel consumption, especially in transport, are setting out new approaches for how climate change mitigation can be achieved.

Full Report. In: Climate Change 2022: Mitigation of Climate Change. Contribution of the WGIII to the AR6 of the IPCC

Doc 3 - IPCC AR6 WGII FR - Page 1352

through involving community members in decision making, increasing the capacity of these communities to respond to climate change (Reid, 2014).

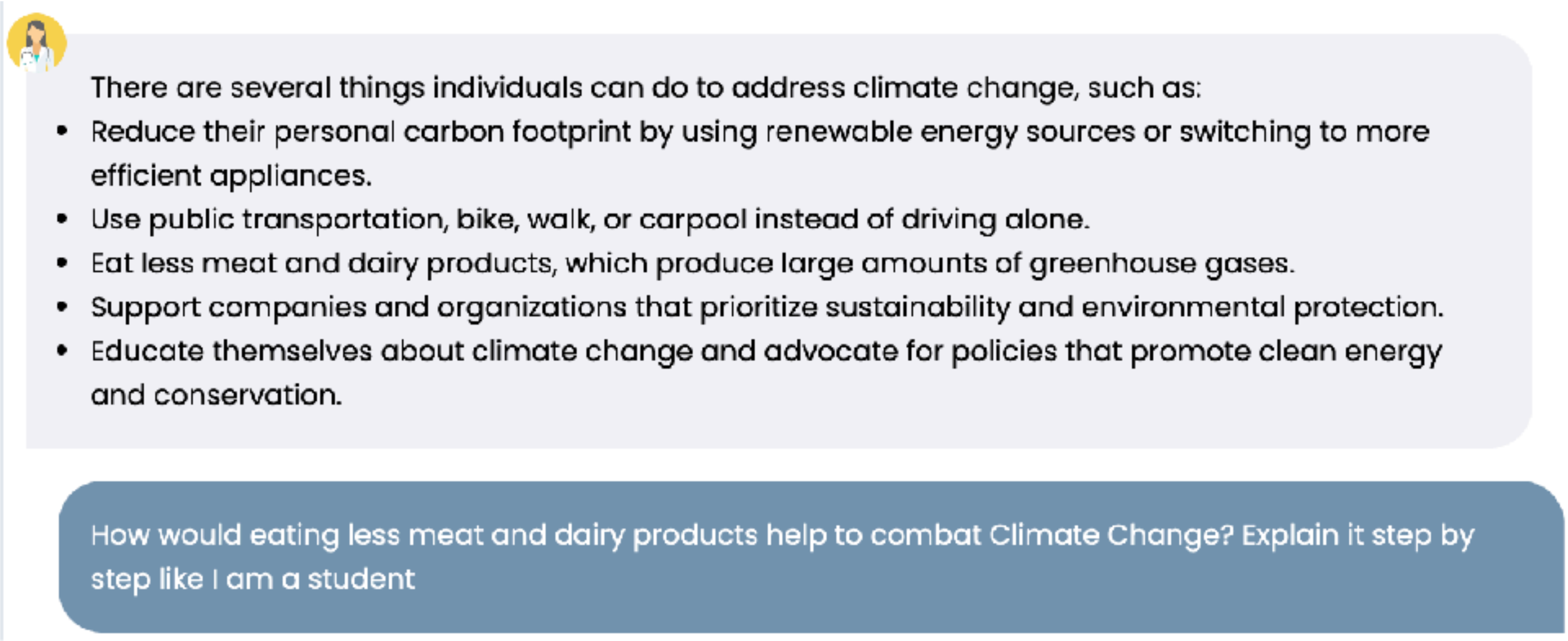
Full Report. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of the WGII to the AR6 of the IPCC

Doc 4 - IPCC AR6 WGIII TS - Page 71

Socio-cultural and lifestyle changes can accelerate climate change mitigation (medium confidence). Among 60 identified actions that could change individual consumption, individual mobility choices have the largest potential to reduce carbon footprints.Prioritising car-free mobility by walking and cycling and adoption of electric mobility could save 2 tCO2-eq cap-1 yr-1. Other options with high mitigation potential include reducing air travel, cooling setpoint adjustments, reduced appliance use, shifts to public transit, and shifting consumption towards plant-based diets. {5.3.1, 5.3.1.2, Figure 5.8}

Ask **IPCC**

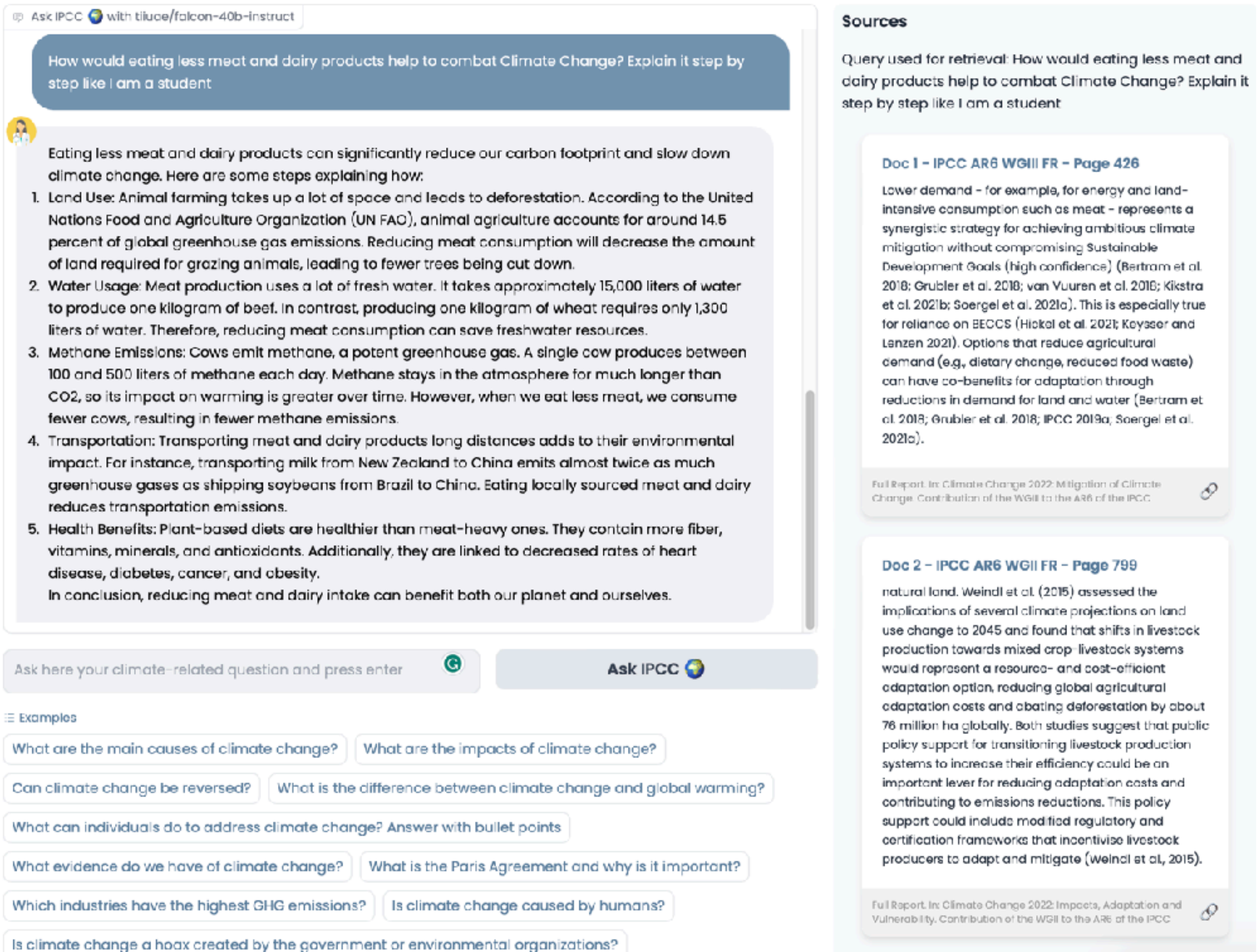
falcon-40b-instruct



There are several things individuals can do to address climate change, such as:

- Reduce their personal carbon footprint by using renewable energy sources or switching to more efficient appliances.
- Use public transportation, bike, walk, or carpool instead of driving alone.
- Eat less meat and dairy products, which produce large amounts of greenhouse gases.
- Support companies and organizations that prioritize sustainability and environmental protection.
- Educate themselves about climate change and advocate for policies that promote clean energy and conservation.

How would eating less meat and dairy products help to combat Climate Change? Explain it step by step like I am a student



How would eating less meat and dairy products help to combat Climate Change? Explain it step by step like I am a student

Eating less meat and dairy products can significantly reduce our carbon footprint and slow down climate change. Here are some steps explaining how:

1. Land Use: Animal farming takes up a lot of space and leads to deforestation. According to the United Nations Food and Agriculture Organization (UN FAO), animal agriculture accounts for around 14.5 percent of global greenhouse gas emissions. Reducing meat consumption will decrease the amount of land required for grazing animals, leading to fewer trees being cut down.
2. Water Usage: Meat production uses a lot of fresh water. It takes approximately 15,000 liters of water to produce one kilogram of beef. In contrast, producing one kilogram of wheat requires only 1,300 liters of water. Therefore, reducing meat consumption can save freshwater resources.
3. Methane Emissions: Cows emit methane, a potent greenhouse gas. A single cow produces between 100 and 500 liters of methane each day. Methane stays in the atmosphere for much longer than CO2, so its impact on warming is greater over time. However, when we eat less meat, we consume fewer cows, resulting in fewer methane emissions.
4. Transportation: Transporting meat and dairy products long distances adds to their environmental impact. For instance, transporting milk from New Zealand to China emits almost twice as much greenhouse gases as shipping soybeans from Brazil to China. Eating locally sourced meat and dairy reduces transportation emissions.
5. Health Benefits: Plant-based diets are healthier than meat-heavy ones. They contain more fiber, vitamins, minerals, and antioxidants. Additionally, they are linked to decreased rates of heart disease, diabetes, cancer, and obesity.

In conclusion, reducing meat and dairy intake can benefit both our planet and ourselves.

Ask here your climate-related question and press enter
Ask IPCC 🌍
Examples
What are the main causes of climate change?
What are the impacts of climate change?
Can climate change be reversed?
What is the difference between climate change and global warming?
What can individuals do to address climate change? Answer with bullet points
What evidence do we have of climate change?
What is the Paris Agreement and why is it important?
Which industries have the highest GHG emissions?
Is climate change caused by humans?
Is climate change a hoax created by the government or environmental organizations?

Sources

Query used for retrieval: How would eating less meat and dairy products help to combat Climate Change? Explain it step by step like I am a student

Doc 1 - IPCC AR6 WGIII FR - Page 426

Lower demand – for example, for energy and land-intensive consumption such as meat – represents a synergistic strategy for achieving ambitious climate mitigation without compromising Sustainable Development Goals (high confidence) (Bertram et al. 2018; Grubler et al. 2018; van Vuuren et al. 2018; Kikstra et al. 2021b; Soergel et al. 2021a). This is especially true for reliance on BECCS (Hickel et al. 2021; Keysser and Lenzen 2021). Options that reduce agricultural demand (e.g., dietary change, reduced food waste) can have co-benefits for adaptation through reductions in demand for land and water (Bertram et al. 2018; Grubler et al. 2018; IPCC 2019a; Soergel et al. 2021a).

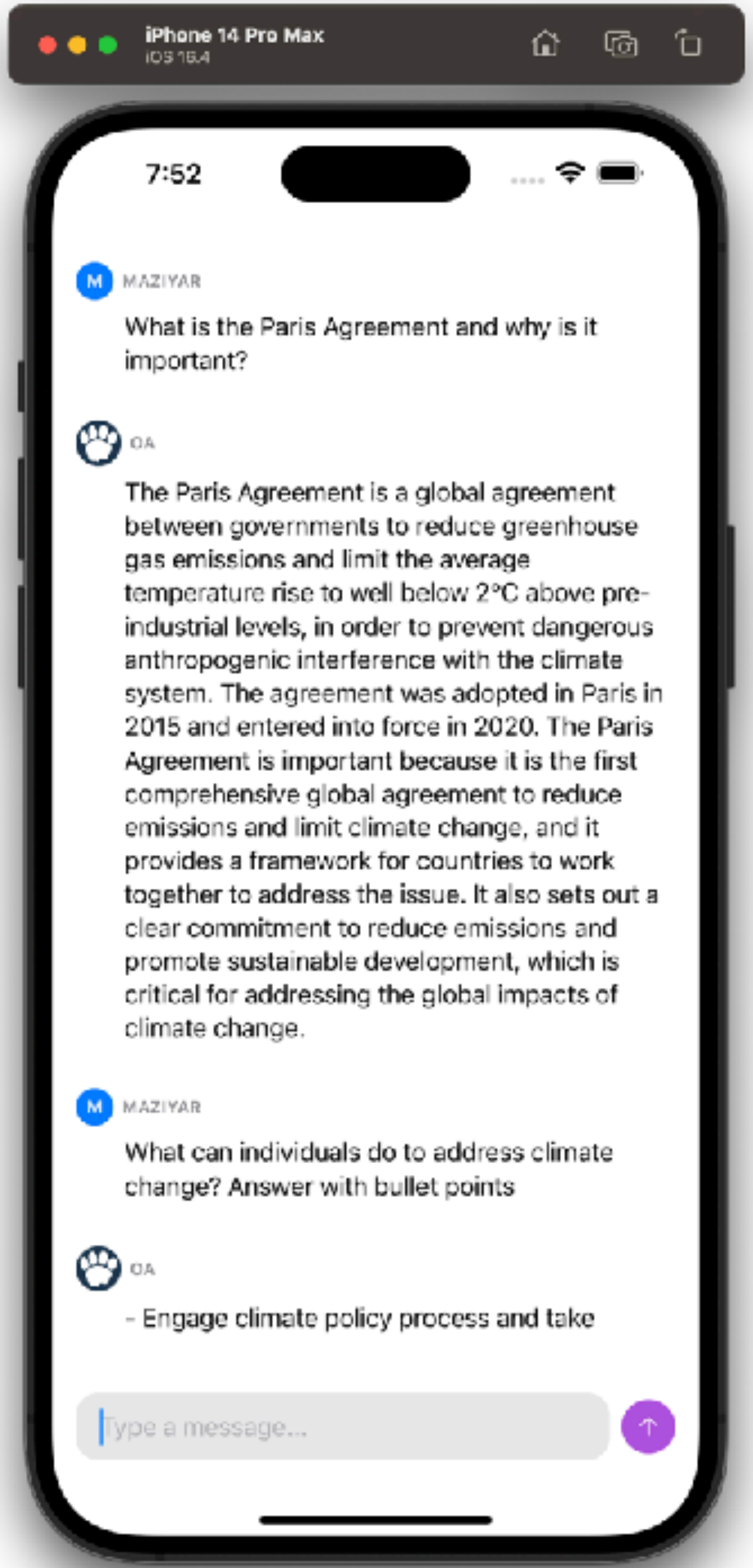
Full Report. In: Climate Change 2022: Mitigation of Climate Change. Contribution of the WGIII to the AR6 of the IPCC

Doc 2 - IPCC AR6 WGII FR - Page 799

natural land. Weindl et al. (2015) assessed the implications of several climate projections on land use change to 2045 and found that shifts in livestock production towards mixed crop-livestock systems would represent a resource- and cost-efficient adaptation option, reducing global agricultural adaptation costs and abating deforestation by about 76 million ha globally. Both studies suggest that public policy support for transitioning livestock production systems to increase their efficiency could be an important lever for reducing adaptation costs and contributing to emissions reductions. This policy support could include modified regulatory and certification frameworks that incentivise livestock producers to adapt and mitigate (Weindl et al., 2015).

Full Report. In: Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of the WGII to the AR6 of the IPCC

Ask **IPCC**
OpenAssistant 12b



M MAZIYAR

What is the Paris Agreement and why is it important?

OA

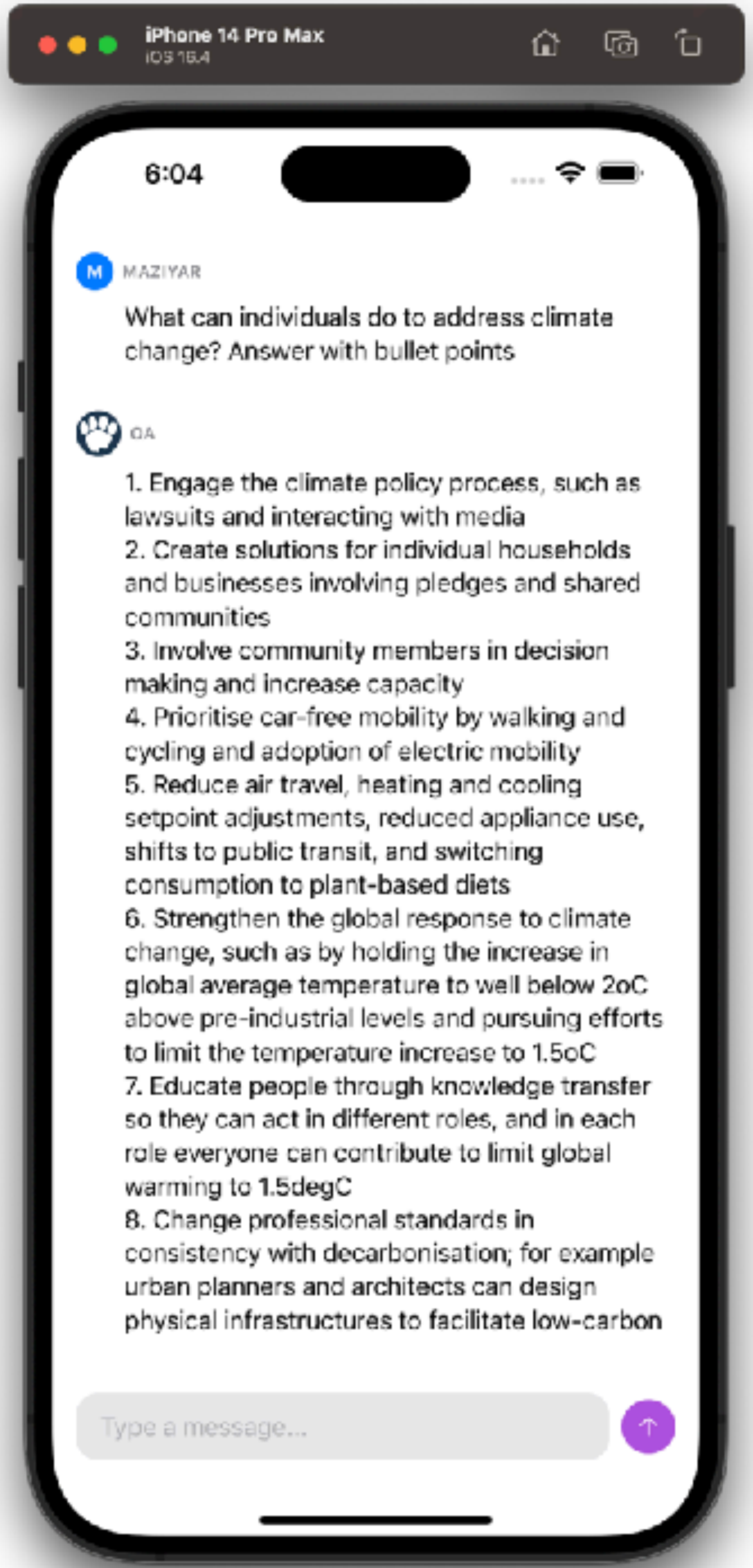
The Paris Agreement is a global agreement between governments to reduce greenhouse gas emissions and limit the average temperature rise to well below 2°C above pre-industrial levels, in order to prevent dangerous anthropogenic interference with the climate system. The agreement was adopted in Paris in 2015 and entered into force in 2020. The Paris Agreement is important because it is the first comprehensive global agreement to reduce emissions and limit climate change, and it provides a framework for countries to work together to address the issue. It also sets out a clear commitment to reduce emissions and promote sustainable development, which is critical for addressing the global impacts of climate change.

M MAZIYAR

What can individuals do to address climate change? Answer with bullet points

OA

- Engage climate policy process and take



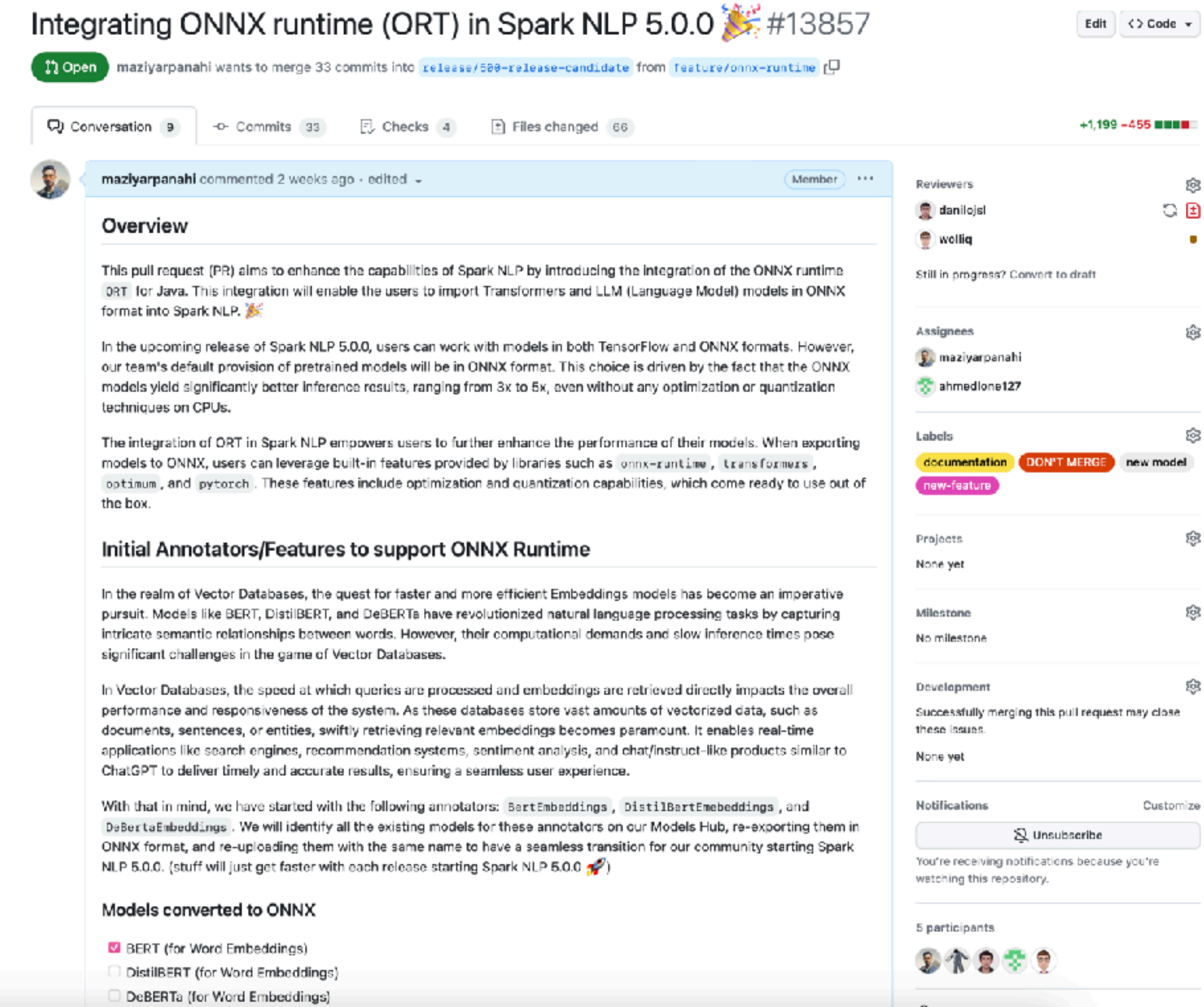
M MAZIYAR

What can individuals do to address climate change? Answer with bullet points

OA

1. Engage the climate policy process, such as lawsuits and interacting with media
2. Create solutions for individual households and businesses involving pledges and shared communities
3. Involve community members in decision making and increase capacity
4. Prioritise car-free mobility by walking and cycling and adoption of electric mobility
5. Reduce air travel, heating and cooling setpoint adjustments, reduced appliance use, shifts to public transit, and switching consumption to plant-based diets
6. Strengthen the global response to climate change, such as by holding the increase in global average temperature to well below 2oC above pre-industrial levels and pursuing efforts to limit the temperature increase to 1.5oC
7. Educate people through knowledge transfer so they can act in different roles, and in each role everyone can contribute to limit global warming to 1.5degC
8. Change professional standards in consistency with decarbonisation; for example urban planners and architects can design physical infrastructures to facilitate low-carbon

cnrs Institut des Systèmes Complexes Paris Île-de-France



Integrating ONNX runtime (ORT) in Spark NLP 5.0.0 🎉 #13857

Open maziyarpanahi wants to merge 33 commits into `release/500-release-candidate` from `feature/onnx-runtime`

maziyarpanahi commented 2 weeks ago · edited

Overview

This pull request (PR) aims to enhance the capabilities of Spark NLP by introducing the integration of the ONNX runtime `ORT` for Java. This integration will enable the users to import Transformers and LLM (Language Model) models in ONNX format into Spark NLP. 🎉

In the upcoming release of Spark NLP 5.0.0, users can work with models in both TensorFlow and ONNX formats. However, our team's default provision of pretrained models will be in ONNX format. This choice is driven by the fact that the ONNX models yield significantly better inference results, ranging from 3x to 5x, even without any optimization or quantization techniques on CPUs.

The integration of ORT in Spark NLP empowers users to further enhance the performance of their models. When exporting models to ONNX, users can leverage built-in features provided by libraries such as `onnx-runtime`, `transformers`, `optimum`, and `pytorch`. These features include optimization and quantization capabilities, which come ready to use out of the box.

Initial Annotators/Features to support ONNX Runtime

In the realm of Vector Databases, the quest for faster and more efficient Embeddings models has become an imperative pursuit. Models like BERT, DistilBERT, and DeBERTa have revolutionized natural language processing tasks by capturing intricate semantic relationships between words. However, their computational demands and slow inference times pose significant challenges in the game of Vector Databases.

In Vector Databases, the speed at which queries are processed and embeddings are retrieved directly impacts the overall performance and responsiveness of the system. As these databases store vast amounts of vectorized data, such as documents, sentences, or entities, swiftly retrieving relevant embeddings becomes paramount. It enables real-time applications like search engines, recommendation systems, sentiment analysis, and chat/instruct-like products similar to ChatGPT to deliver timely and accurate results, ensuring a seamless user experience.

With that in mind, we have started with the following annotators: `BertEmbeddings`, `DistilBertEmbeddings`, and `DeBertaEmbeddings`. We will identify all the existing models for these annotators on our Models Hub, re-exporting them in ONNX format, and re-uploading them with the same name to have a seamless transition for our community starting Spark NLP 5.0.0. (stuff will just get faster with each release starting Spark NLP 5.0.0 🚀)

Models converted to ONNX

- ☑ BERT (for Word Embeddings)
- ☐ DistilBERT (for Word Embeddings)
- ☐ DeBERTa (for Word Embeddings)

Reviewers

danilojsl

wolliq

Still in progress? Convert to draft

Assignees

maziyarpanahi

ahmedlone127

Labels

documentation DON'T MERGE new model new-feature

Projects

None yet

Milestone

No milestone

Development

Successfully merging this pull request may close these issues.

None yet

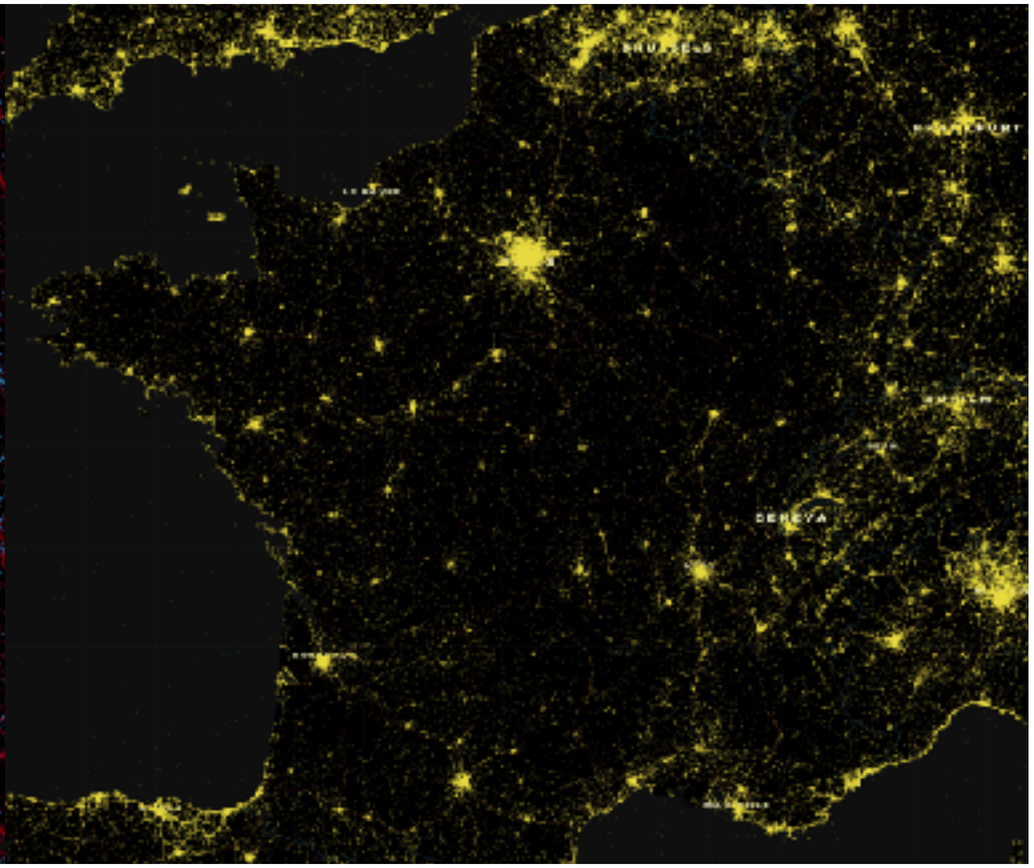
Notifications

You're receiving notifications because you're watching this repository.

5 participants

DES RÉSEAUX SOCIAUX DANS NOS URNES
POLITOSCOPE.ORG #PRESIDENTIELLE2017

JOURNALIST PLATFORM
WHERE JOURNALISM MEETS BIG DATA!
Welcome!
IT'S ALL ABOUT DATA JOURNALISM!



Thank You!

Maziyar **PANAHI**

maziyar.panahi@iscpif.fr