Tracing Epidemic Outbreaks on Twitter

Social Network Analysis for Doctors without Borders

Kateřina Lesch

Contents

Aim of the project:

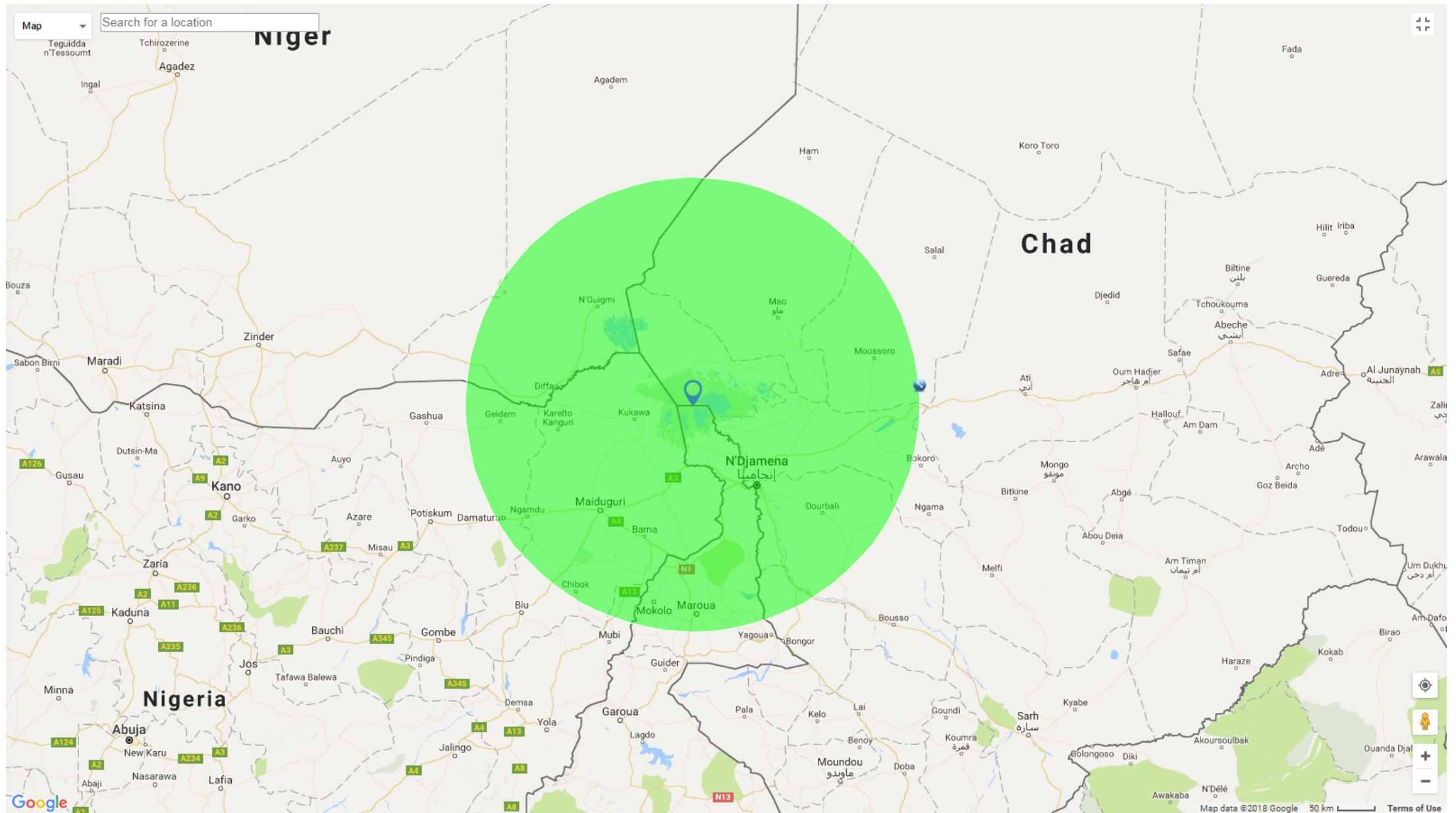monitoring disease outbreaks and epidemic spreading with Twitter data

credits: Jan Böhm

Regional cholera outbreak threat in west Africa as cases increase eight-fold in Lake Chad basin

f share 🐦 ✉

🔖 Save

More than 500 people have died from cholera in the Lake Chad region since the start of the year, representing the worst outbreak to hit the area in four years, the UN said on Wednesday.
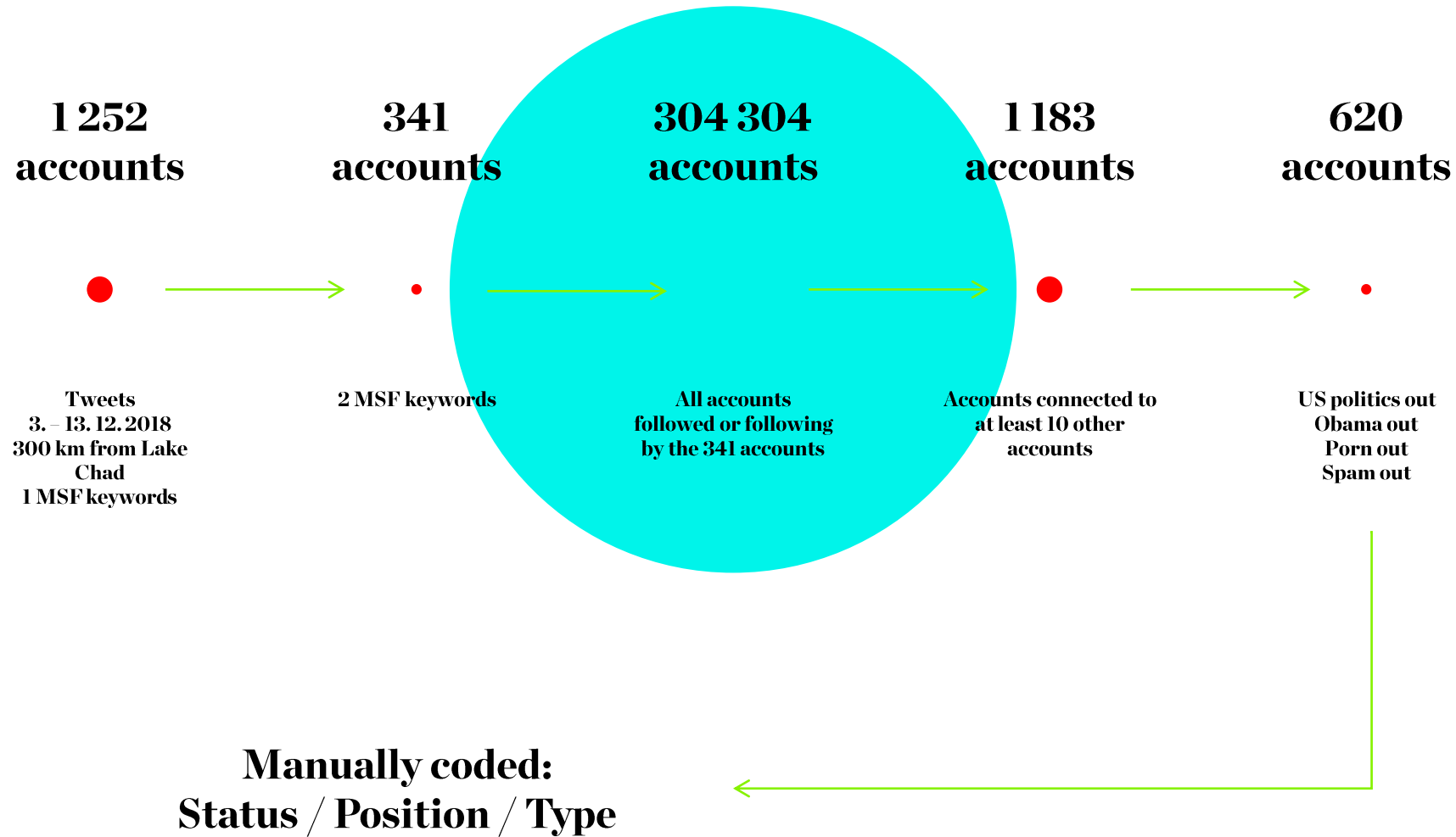

Dozens dead in Lake Chad cholera outbreaks

There are an estimated 1.3 - 4 million cases of cholera worldwide each year
BULAWAYO/REUTERS

Lake Chad, photo by **CAJ News Africa, Chad Bureau**

from *EMEKA OKONKWO* in Abuja, Nigeria
ABUJA, (CAJ News) – AN outbreak of cholera has claimed the lives of at least 39 people in the Lake Chad Basin since the beginning of 2019.

Methodology

1 252
accounts

341
accounts

304 304
accounts

1 183
accounts

620
accounts

Tweets
3. – 13. 12. 2018
300 km from Lake
Chad
1 MSF keywords

2 MSF keywords

All accounts
followed or following
by the 341 accounts

Accounts connected to
at least 10 other
accounts

US politics out
Obama out
Porn out
Spam out

Manually coded:
Status / Position / Type

List of keywords

English	French	Arabic	English	French	Arabic
Access	accès	وصول	Insurgency	insurrection	تمرُد
Aerial	aérien	جوي	Killed	mort	قتل
Aid	aide humanitaire	إغاثة	Lake Chad	lac Tchad	بحيرة تشاد
Aid worker	travailleur humanitaire	عامل إغاثة	Malaria	Palud	ماتريا
Army	armée	جيش		Malnutrition	سوء التغذية
Assistance	assistance	مساعدة	Mama Nur		
Attack	attaque	هجوم	Medical	médical	طبي
Barnawi		برناوي	Military	militaire	عسكري
Boko		بوكو	Negociations	négociations	مفاوضات
Boko Haram		بوكو حرام	Nigeriacrisis	Nigeriacrisis	الأزمة النيجيرية
Bomb	bombe	قنبلة/قذيفة		Nutrition	تغذية
Bombardment	bombardement	القصف	Outbreak	épidémie	وباء
Chibok	chibok	شيبوك	Refugees	réfugiés	لاجئين
	enfants	أطفال		sexuel	جنسي
Cholera		كوليرا	Shekau	shekau	شيكاو
Conduct of war	conduite de la guerre	معايير الحرب	Soldier	soldat	جندي
Conflict	conflit	نزاع	Starving	affamé	تُجُوِع أو تضور جوعًا
Crisis	crise	أزمة	Suicide Bomb	attentat suicide	تفجير انتحاري
Delivery	livraison	ولادة	Terrorism	terrorisme	إرهاب
Displaced	déplacés	نازح	Vaccines	vaccins	لقاحات
Displacement	déplacement	نزوح	Violence	violences	عنف
Doctors	médecins	أطباء	War	guerre	حرب
Emergency	urgence	طارئ	Water	eau	ماء
Explosion		إنفجار	Terrosim		إرهاب
Food	alimentation	غذاء	National Politics	politique nationale	السياسات الوطنية
Haram	haram	حرام	Military actions	actions militaires	عمليات عسكرية
Health	santé	صحة	Kidnappings	kidnapping	اختطاف
Hep E	Hep E	التهاب الكبد E	Famine		مجاعة
Hepatitis E	hépatite E	التهاب الكبد E	Developmend	développement	تنمية
Humanitarian	humanitaire	إنساني	Armed Forces	forcées armées	القوات المسلحة
Hungry	faim	جائع	National Police	police nationale	الشرطة الوطنية أو الشرطة المحلية
Hygiene	hygiène	الصحة العامة		Kamikazes	كاميكازي
IDP's	PDI	نازحون		Attentat	انتهاك
Injured	blessés	مصاب			

Relationships between entities

credits: Matouš Pilnáček
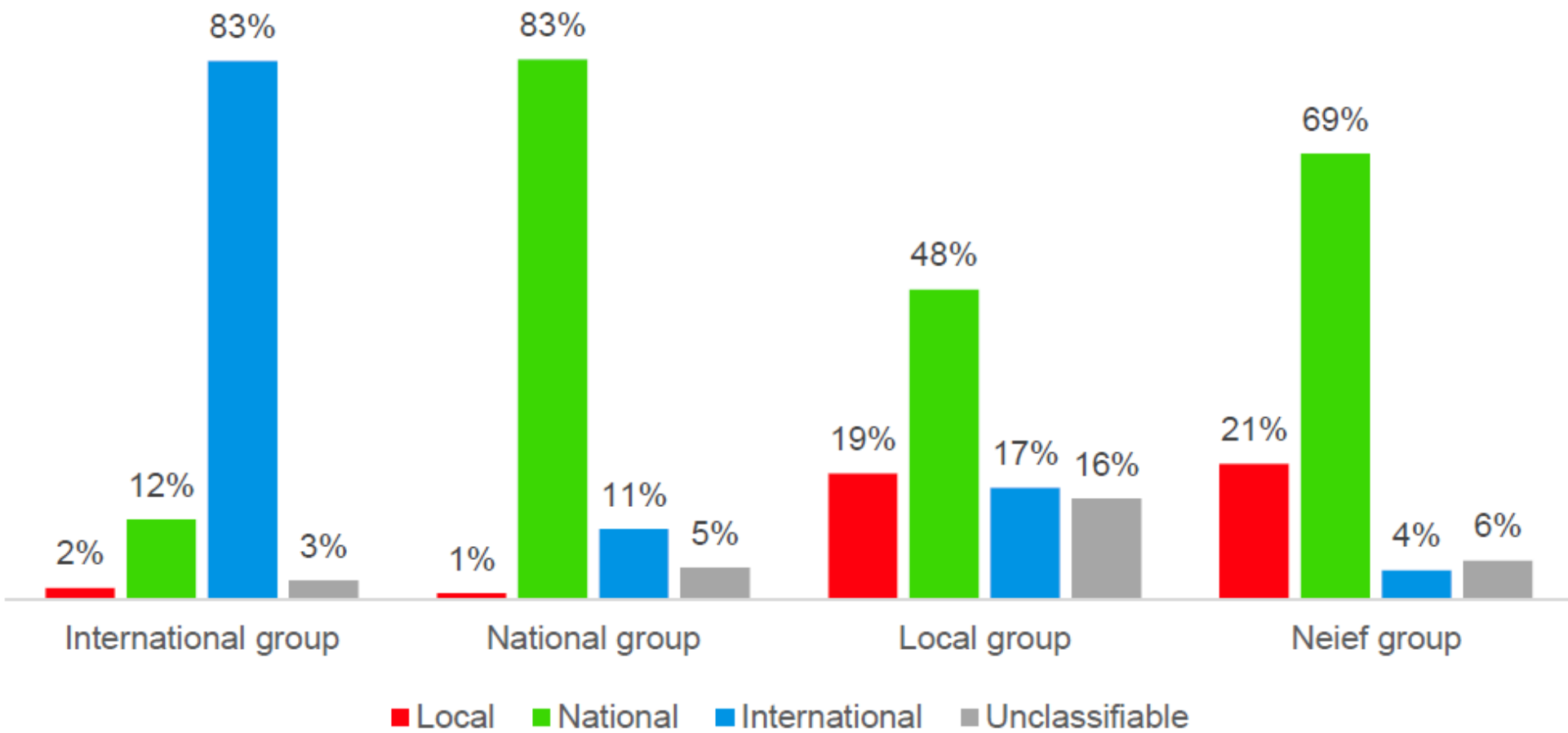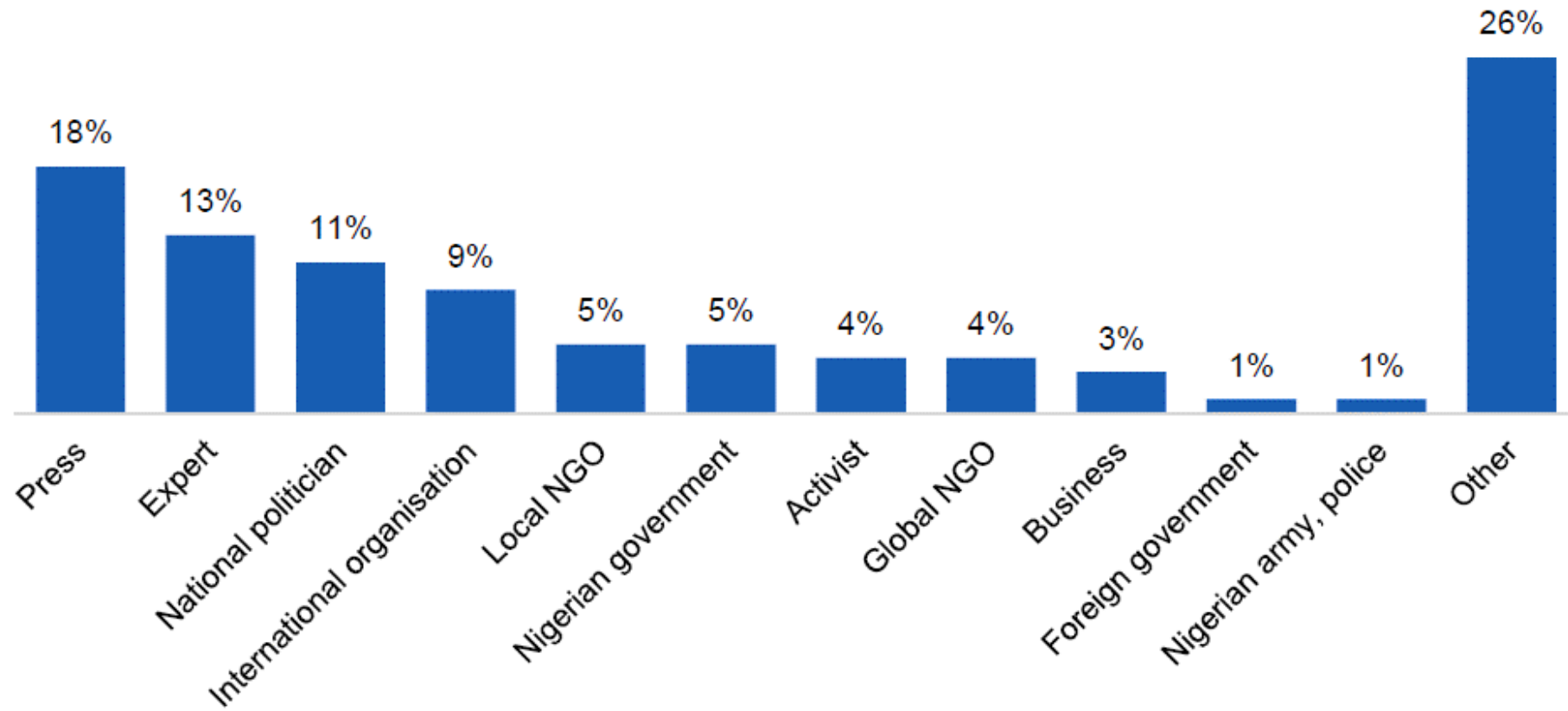
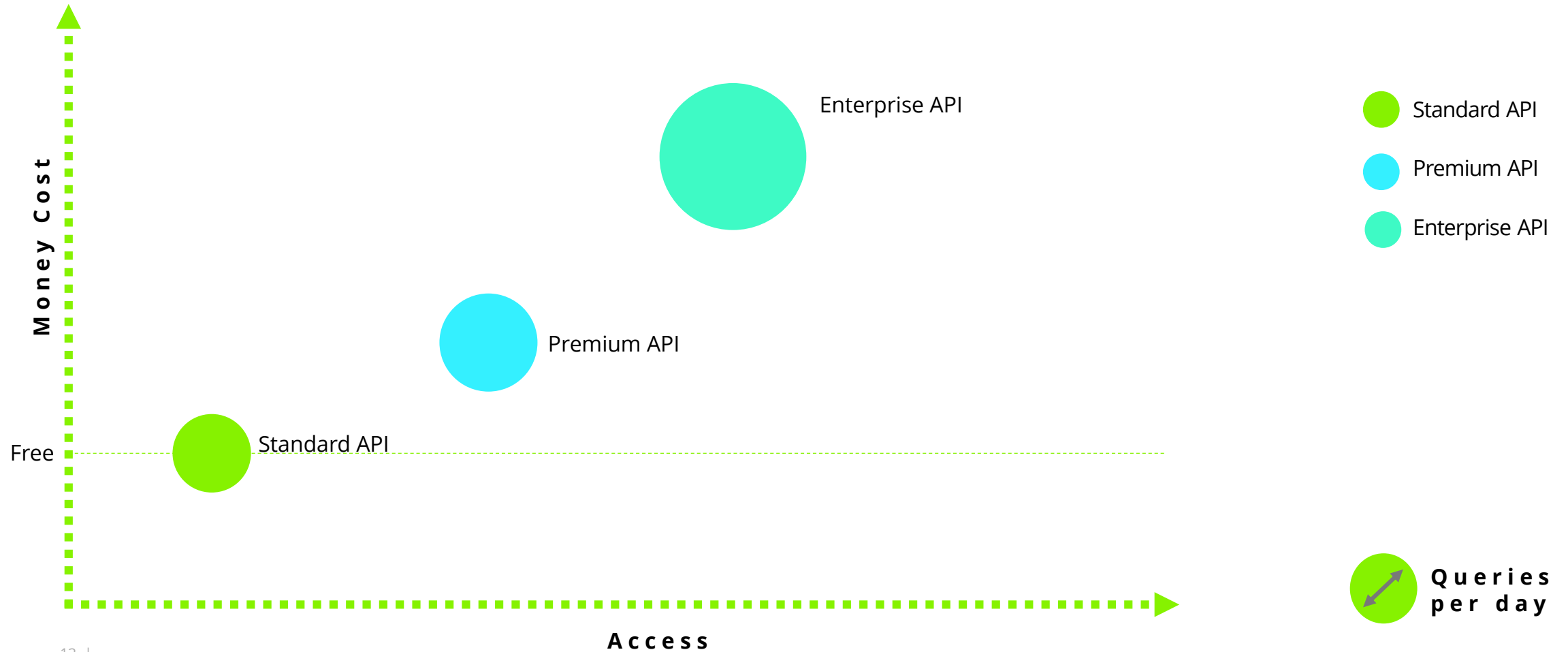# Twitter API

Two ways to obtain tweets from Twitter

Twitter API with the usage of Twitter Developer's account

Legal

Scraping the Twitter data with creating query by building URL
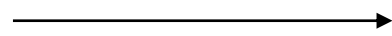
Illegal

Twitter API accounts

- Standard API
- Premium API
- Enterprise API

Queries per day

Twitter API accounts

Enterprise API

??? Tweets per request

Twitter does not provide price for enterprise accounts. This data are available if requested

Solution ⟶ Download tweets from on-going streaming
(max **2000** accounts
500 keywords)

Tools for Twitter monitoring

Twitter API
PROVIDES ACCESS TO TWITTER

Python
DOWNLOADS TWEETS AND MAKES PREPROCESSING

Relativity
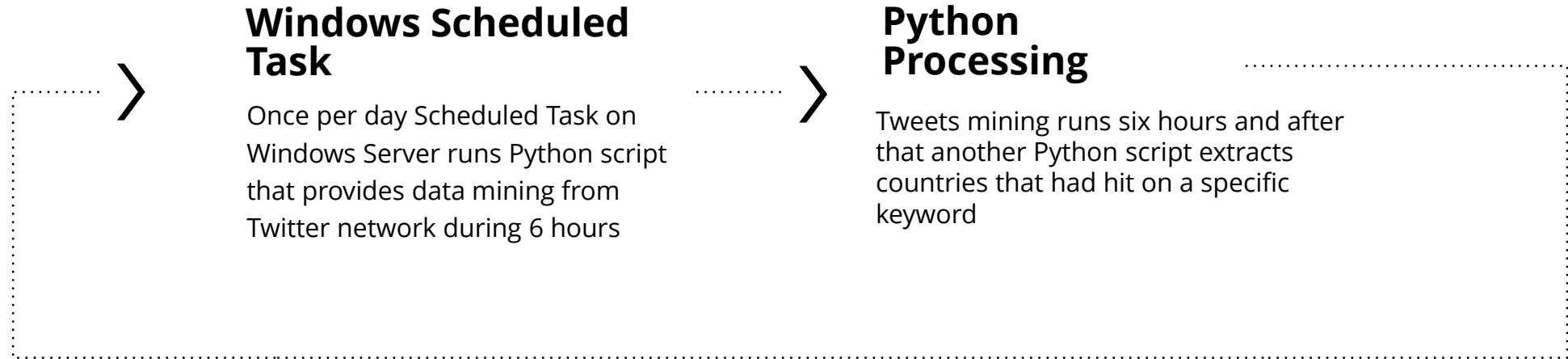**CREATES SEARCH INDEXES AND LOOKUPS THE KEY TERMS**

Brainspace
VISUALIZATION AND SEMANTIC ANALYSIS

Data Processing

Process Set Up (First Solution)

Windows Scheduled Task

Once per day Scheduled Task on Windows Server runs Python script that provides data mining from Twitter network during 6 hours

Python Processing

Tweets mining runs six hours and after that another Python script extracts countries that had hit on a specific keyword

For the POC purposes this process ran approximately 4 days

Twitter API and Python input
(First Solution)

Predefined Data

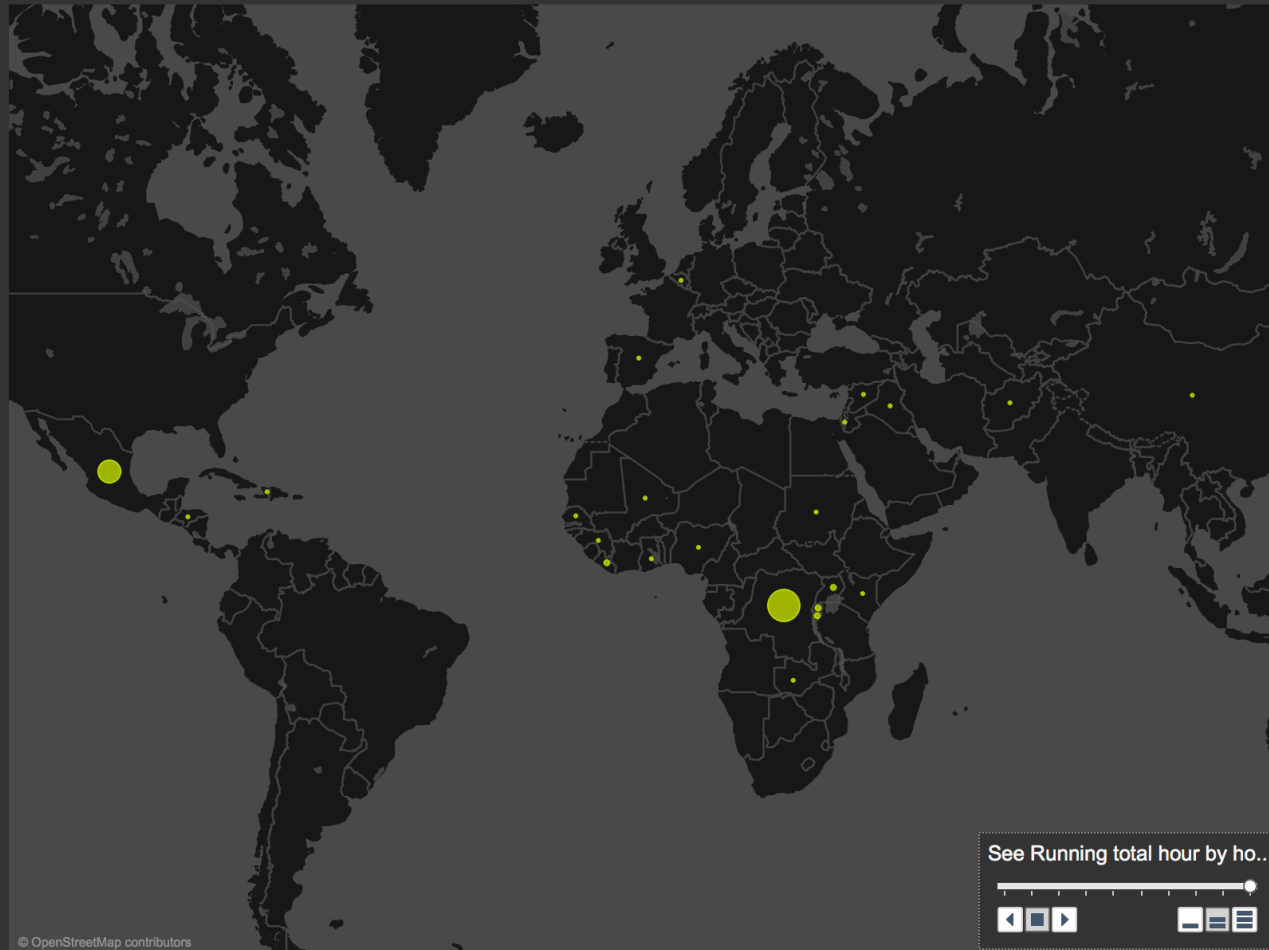We were asked to download tweets with respect to the list of keywords that was previously given

Modified Data

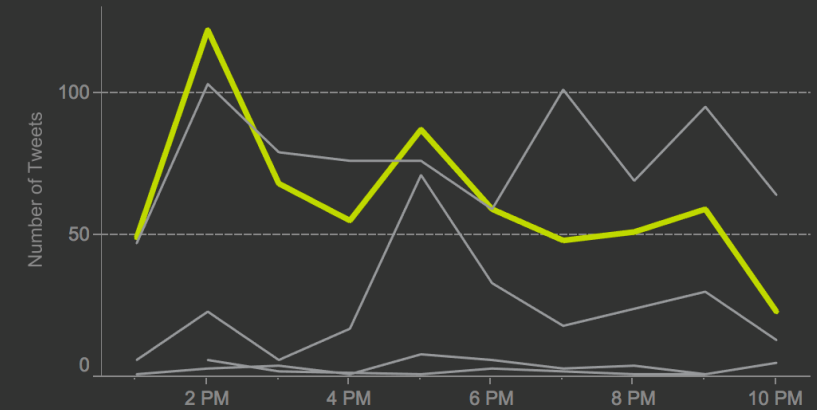Keywords were filtered due to high noise. Filtering reduced amount of False Positive documents
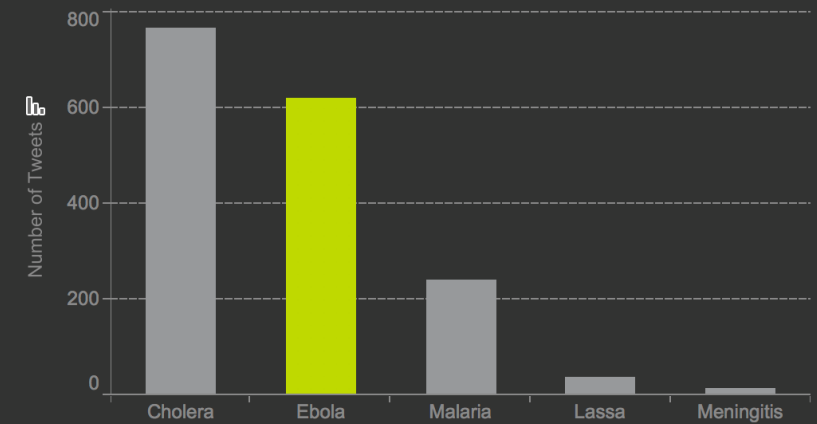
Process Set Up (Second Solution)

Windows Scheduled Task

Each two hours Scheduled Task on Windows Server runs Python script that provides data mining from Twitter network

Python Preprocessing

Tweets mining runs 1 hour and 35 minutes. As a result, we have 25 minutes before next Scheduled Task run to refresh Twitter API limits
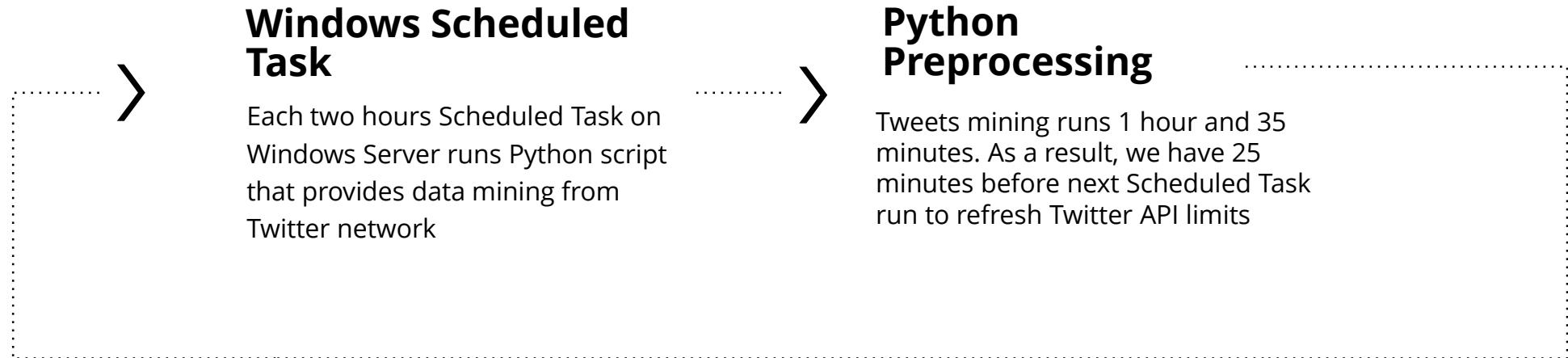
For the POC purposes this process ran approximately 10 days

Twitter API and Python input
(Second Solution)

Predefined Data

Tweets were downloaded with respect to the list of users that was previously given

Modified Data

Twitter API can filter tweets on the basis of User IDs. Thus we had to find all the User IDs according to the list of Usernames

Data Processing

GIVEN DATA

MODIFIED DATA

USERNAME	USER ID

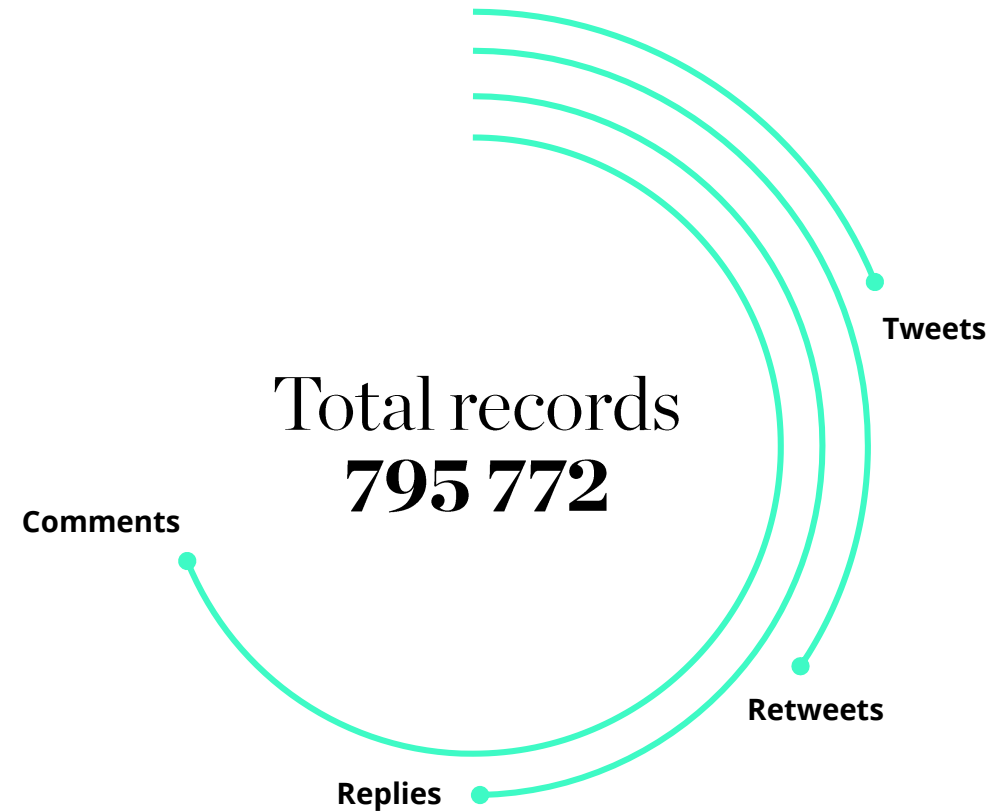|	
M B u h a r i	2 9 3 6 7 1 4 8 4 8
P r o f O s i n b a j o	2 9 1 4 4 4 2 8 7 3
N G R P r e s i d e n t	2 9 6 4 9 5 0 3 1 3
O m o j u w a	1 1 5 6 2 8 2 2 4

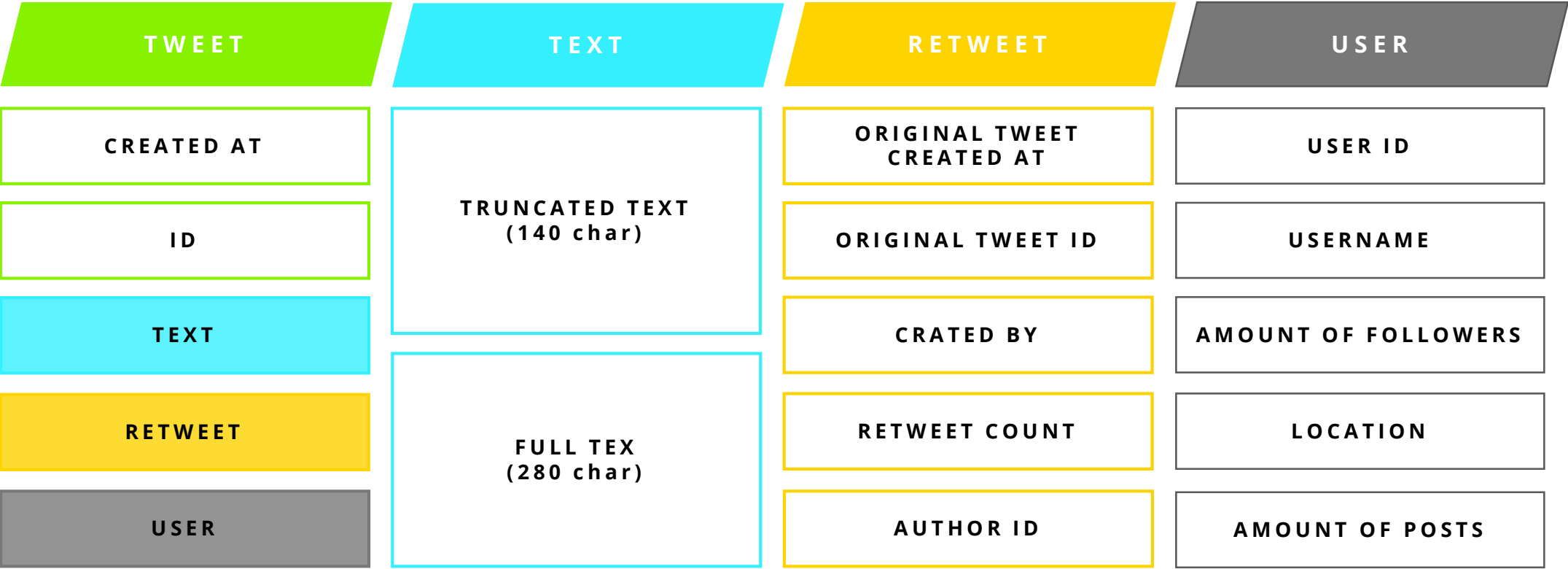# Collected Data

After approximately 10 days of data
collection we obtained

- Tweets

- Retweets

- Comments

- Replies

Total records
795 772

Tweets

Comments

Retweets

Replies

Tweet Structure

TWEET	TEXT	RETWEET	USER
CREATED AT		ORIGINAL TWEET CREATED AT	USER ID
ID	TRUNCATED TEXT (140 char)	ORIGINAL TWEET ID	USERNAME
TEXT		CRATED BY	AMOUNT OF FOLLOWERS
RETWEET	FULL TEX (280 char)	RETWEET COUNT	LOCATION
USER		AUTHOR ID	AMOUNT OF POSTS

Relativity and Brainspace Tandem

Search efficiency

Analytics

Visualization

Relativity®

Relativity is a forensic tool for reviewing legal documents. However, we decided to use it for MSF task purposes

DtSearch index that allows to make fast and efficient key words filtering
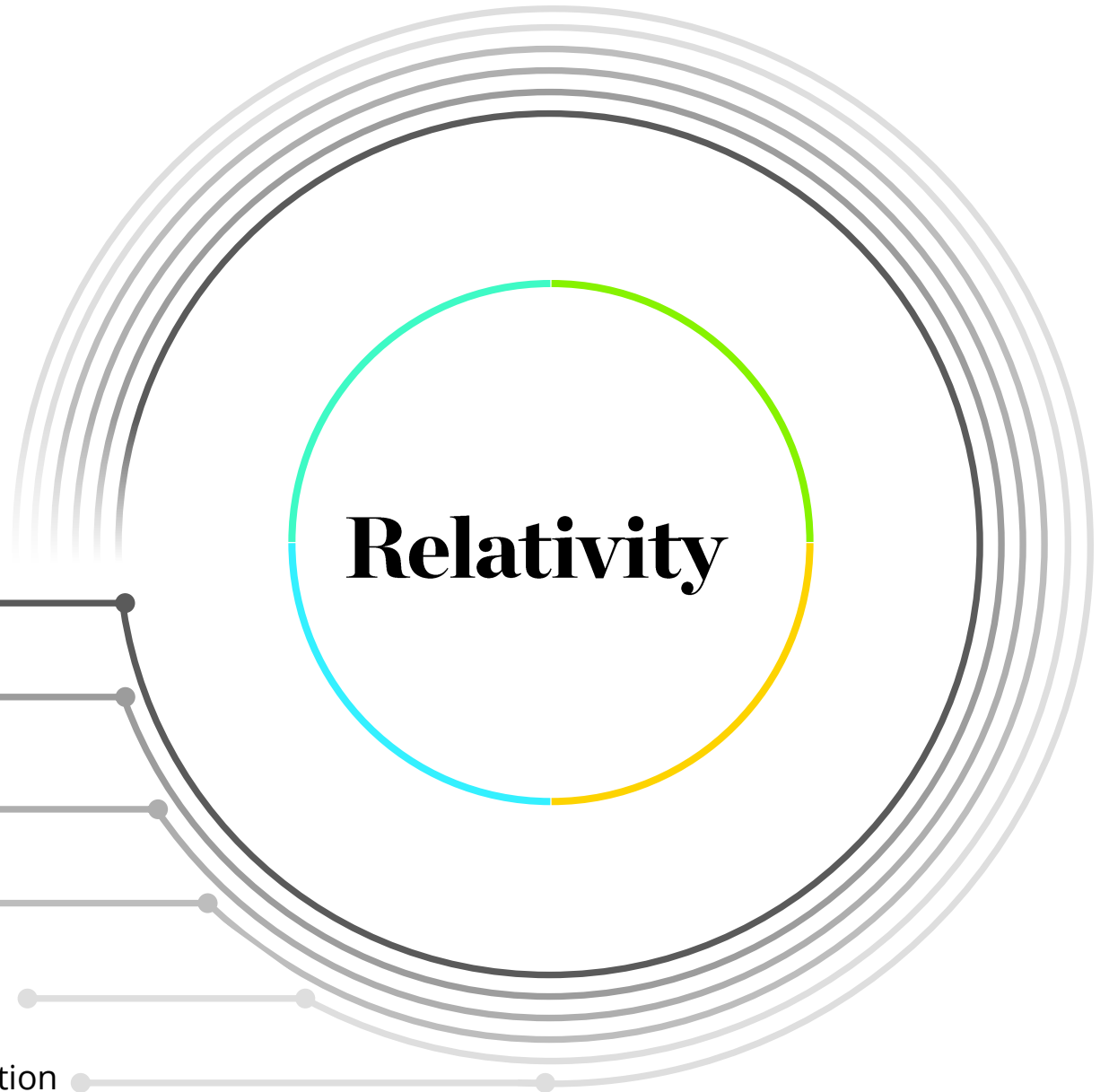
User Interface that displays all needed metadata

Possibility to set up personal view

Create saved searches with many possible conditions

Connection to Brainspace

Visualization

Relativity

Brainspace™

Brainspace is a forensic tool for processing legal documents. Main difference between Relativity and Brainspace is that Brainspace uses AI algorithms to perform dependencies between text and documents

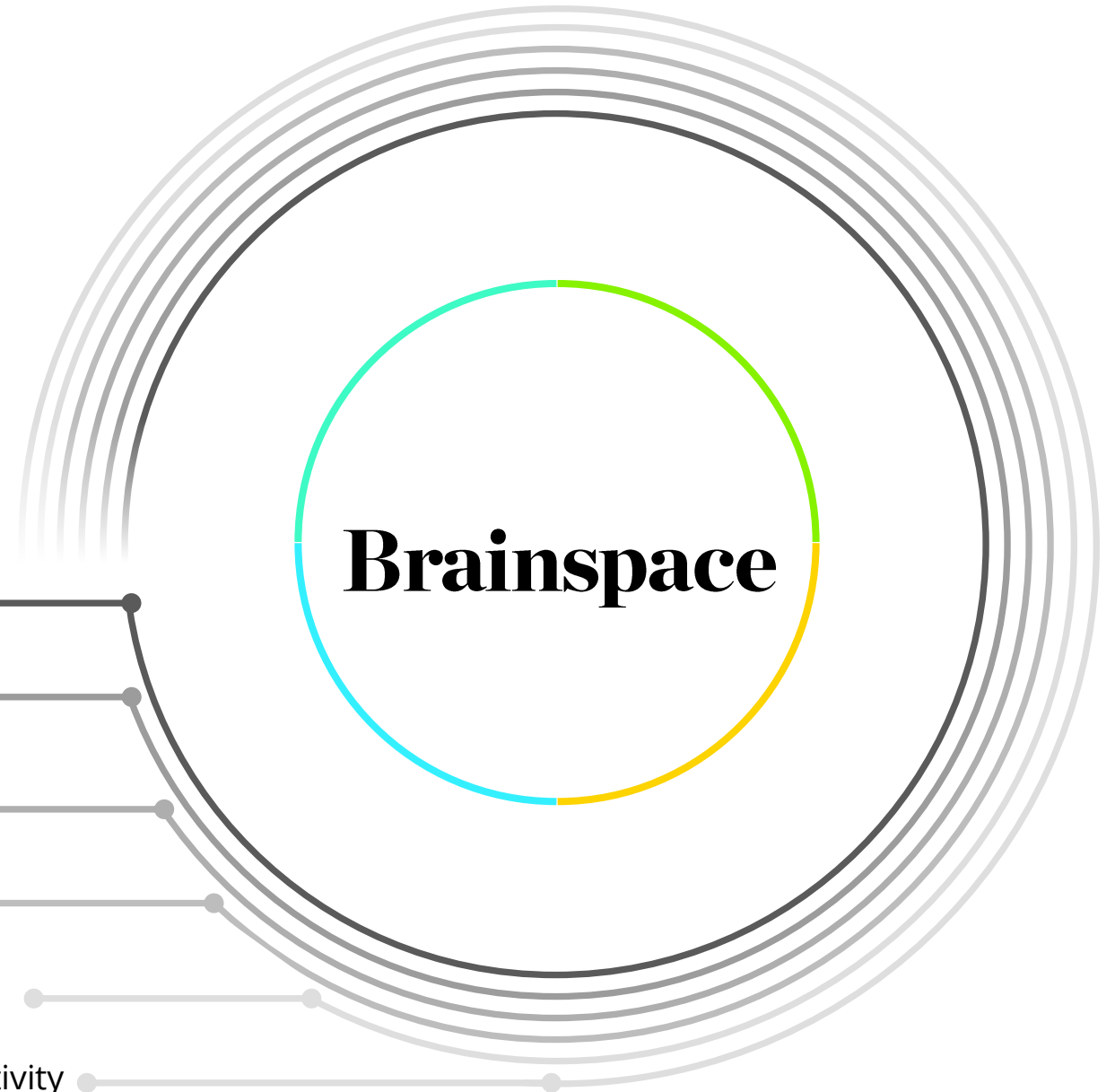Context understanding and identification of similar documents

Advanced searches based on context

Anomalies detection

Relevancy distribution for text documents

Visualization

Connection to Relativity

Brainspace

Discussion?

Lets have a look at demo

Thank you.

Kateřina Lesch (Veselovská)

katerina.lesch@gmail.com

Twitter API accounts

Standard API

100 Tweets per request
15 Requests (max) per 15 minutes
450 Tweets per 15 minutes window
(Tweets from last 7 days)

Premium API

500 Tweets per request
(Full history)

⟶

	Total Requests PER MONTH ❓	Month-to-month PRICE PER MONTH ❓
Paid		
	Up to 100	$99.00
	Up to 250	$224.00
	Up to 500	$399.00
	Up to 1000	$774.00
	Up to 2,500	$1,899.00