How to design a **collection strategy** for monitoring **disinformation** and **hate speech** in **Telegram**

Elisa Muratore [1,2], Thomas Louf [1], Elisa Leonardelli [1,2], Riccardo Gallotti [1,2]

1. **Bruno Kessler Foundation**, Trento Italy   2. **Hatedemics consortium**, CERV-funded project

FONDAZIONE BRUNO KESSLER

Hatedemics

CHuB Complex Human Behaviour Lab

Hatedemics

Fact-Checkers
Misinformation countering + Knowledge preparation

NGO-Operators
Hate countering + Social intervention
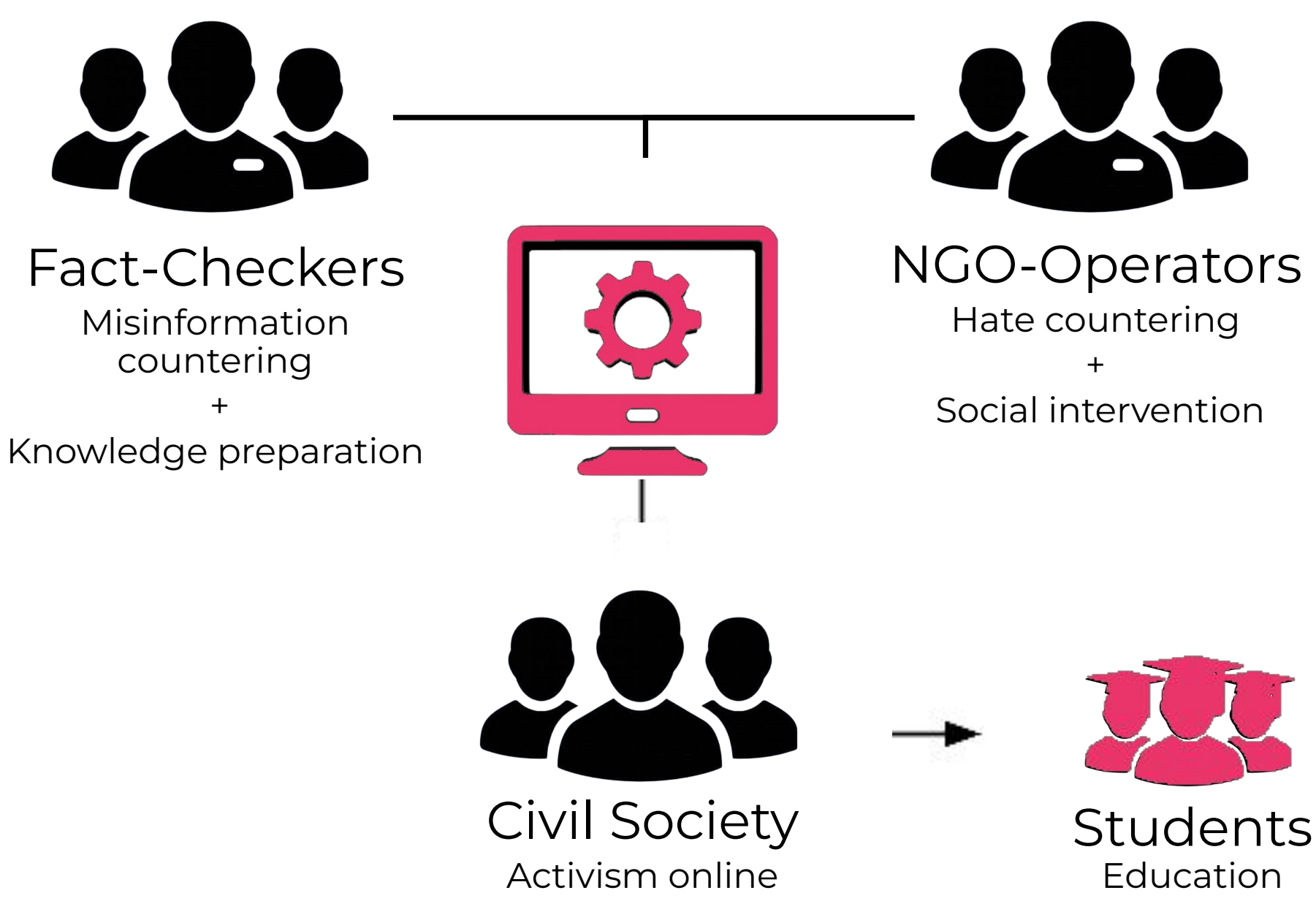
Civil Society
Activism online

Students
Education

Objective

Problem
Traditional social media **data access is restricted,** hindering research on disinformation and hate speech.
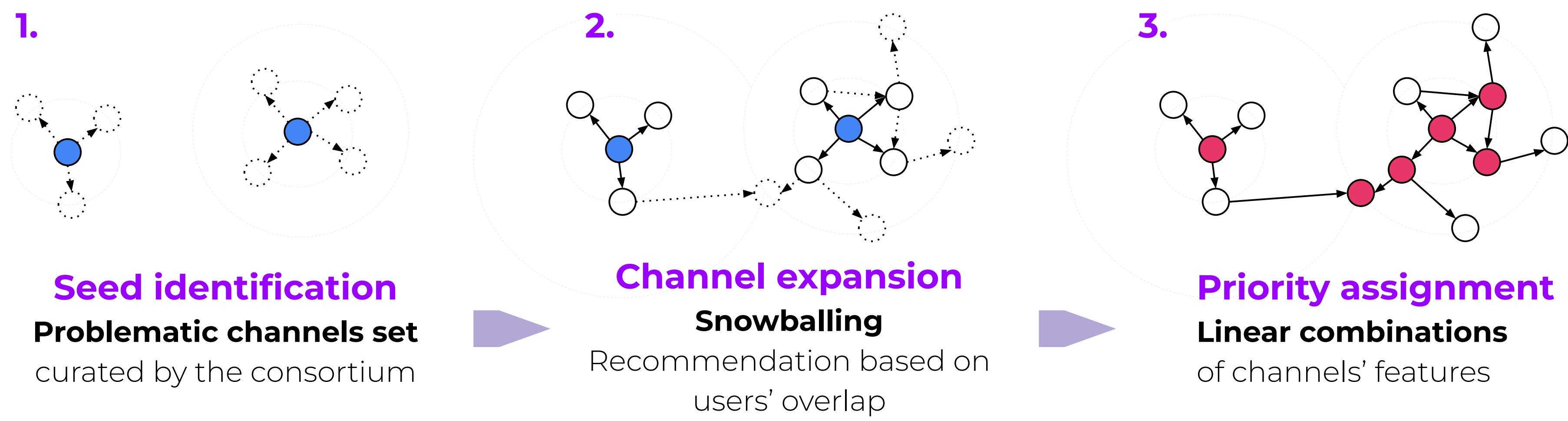
Challenge
Telegram's fragmented structure and lack of keyword search make data collection difficult.

Solution
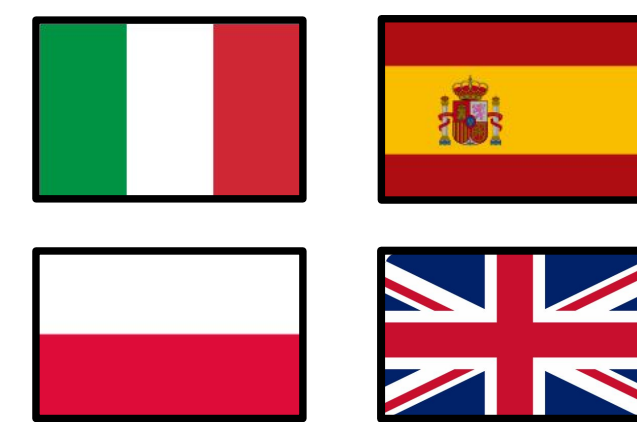We propose a systematic, task-driven **strategy to effectively collect data** from problematic Telegram channels.
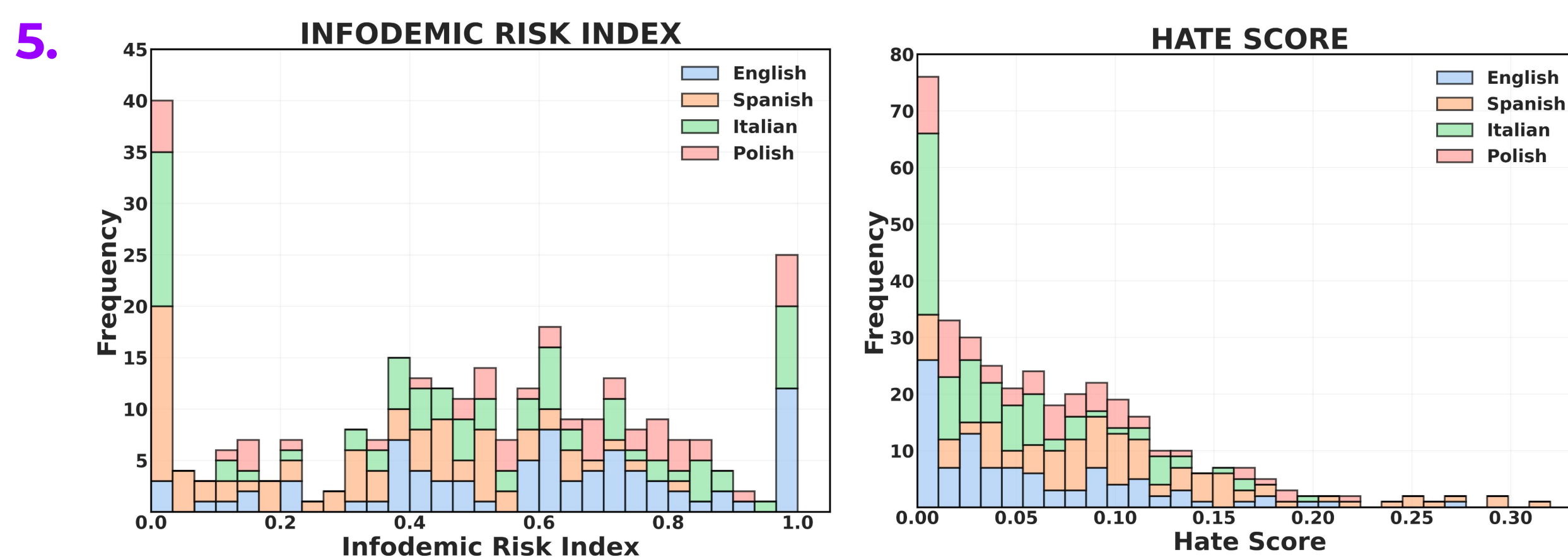
Collection strategy

1. Seed identification
Problematic channels set
curated by the consortium

2. Channel expansion
Snowballing
Recommendation based on users' overlap

3. Priority assignment
Linear combinations
of channels' features

4.

Date	Message	HS	IRI
September 12, 2022 at 10:57 AM	🖼 *Studies show that drinking bleach can cure COVID. Many hospitals are hiding this information to keep their profits up*	❌	✅
September 17, 2022 at 05:50 PM	*All [minority group] are criminals and should be banned from our country. They ruin everything for decent people.*	✅	❌
September 20, 2022 at 02:04 AM	*Hey everyone, just a reminder to stay safe and follow local health guidelines during the pandemic.*	❌	❌
November 15, 2022 at 07:05 PM	*The government is secretly allowing [minority group] to take over our neighborhoods. They are responsible for all the crime and should be expelled immediately.*	✅	✅

Message retrieval
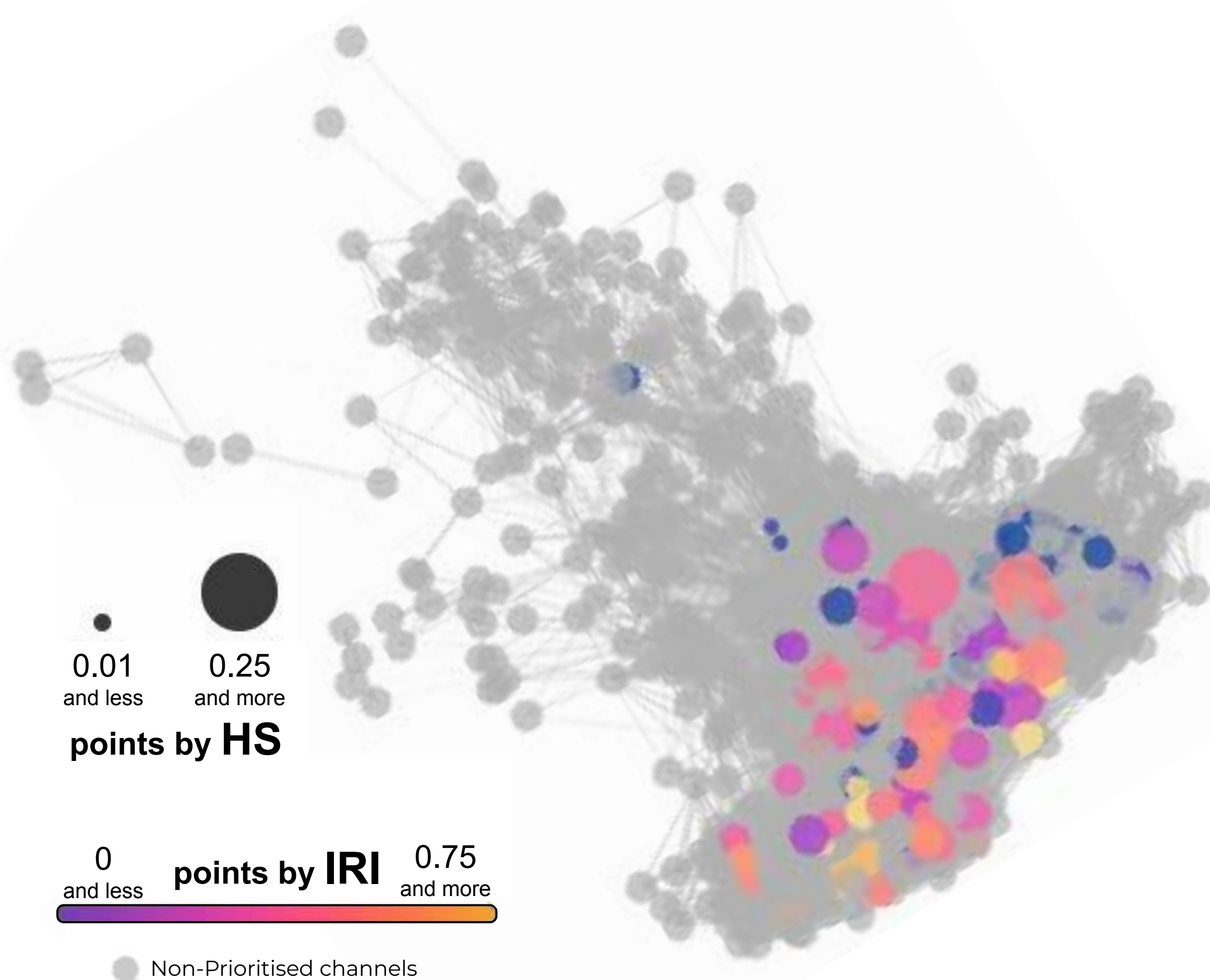**Top-ranked channels**
Over 10M messages
Around 10k channels

Overview

5.

INFODEMIC RISK INDEX — Frequency vs Infodemic Risk Index (English, Spanish, Italian, Polish)

HATE SCORE — Frequency vs Hate Score (English, Spanish, Italian, Polish)

Disinformation and hate speech scores
Infodemic Risk Index (IRI) and Hate Score (HS)

Results

Network

points by **HS**
0.01 and less   0.25 and more

points by **IRI**
0 and less   0.75 and more

Non-Prioritised channels

Correlation

	IRI	HS	CW	NS
Infodemic Risk Index (IRI)	1.00	-0.02	-0.05	-0.12
Hate Score (HS)	-0.02	1.00	-0.43	0.21
Check-worthiness (CW)	-0.05	-0.43	1.00	0.42
Negative Sentiment (NS)	-0.12	0.21	0.42	1.00
Distance from seed	0.21	-0.28	0.20	-0.08
N in recommendation	-0.09	0.17	0.03	0.23
N messages	0.11	0.19	-0.15	0.15
N users	0.04	-0.01	0.05	0.14
Degree centrality	-0.17	0.28	-0.06	0.19
Betweenness centrality	-0.12	-0.05	0.12	0.09
Closeness centrality	-0.20	0.23	-0.07	0.18
Eigenvector centrality	0.06	0.09	0.22	0.12

Pearson correlation
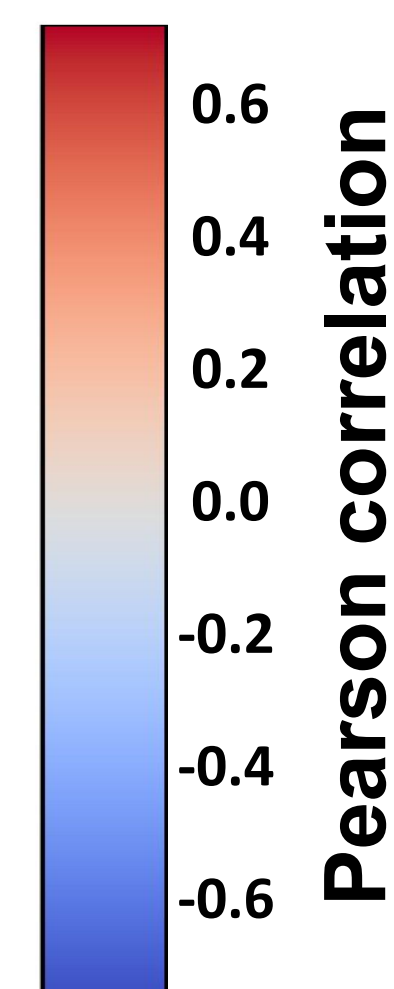
Conclusions

- The proposed strategy is a **scalable, effective** method for collecting data on disinformation and hate speech.

- Collected channels have relatively **high IRI** and **HS**.

IRI: 0.48
HS: 6.71%

Twitter World Average / Hatedemics Average — Infodemic Risk Index
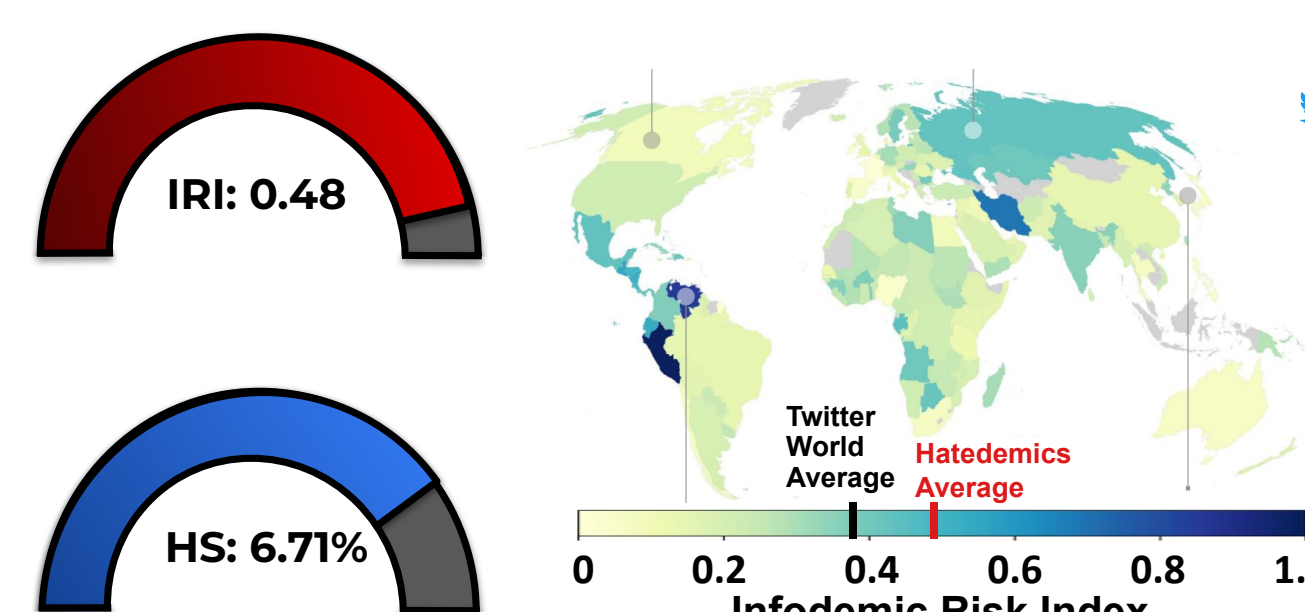
- **Next steps** include comparing alternative priority score methods.