MIDI
NNOVANT

UMRSU

Détection d'événements à partir d'un flux de tweets

MidInnovant de l'UMRsu

Sophie Baillargeon

25 janvier 2018

Personnes impliquées

Partenariat THALES - UNIVERSITÉ LAVAL

Simon Hallé : Spécialiste et Manager du Portfolio Big Data, Thales Canada Recherche et Technologie (TRT).

Sophie Baillargeon : Moi, étudiante au doctorat en statistique

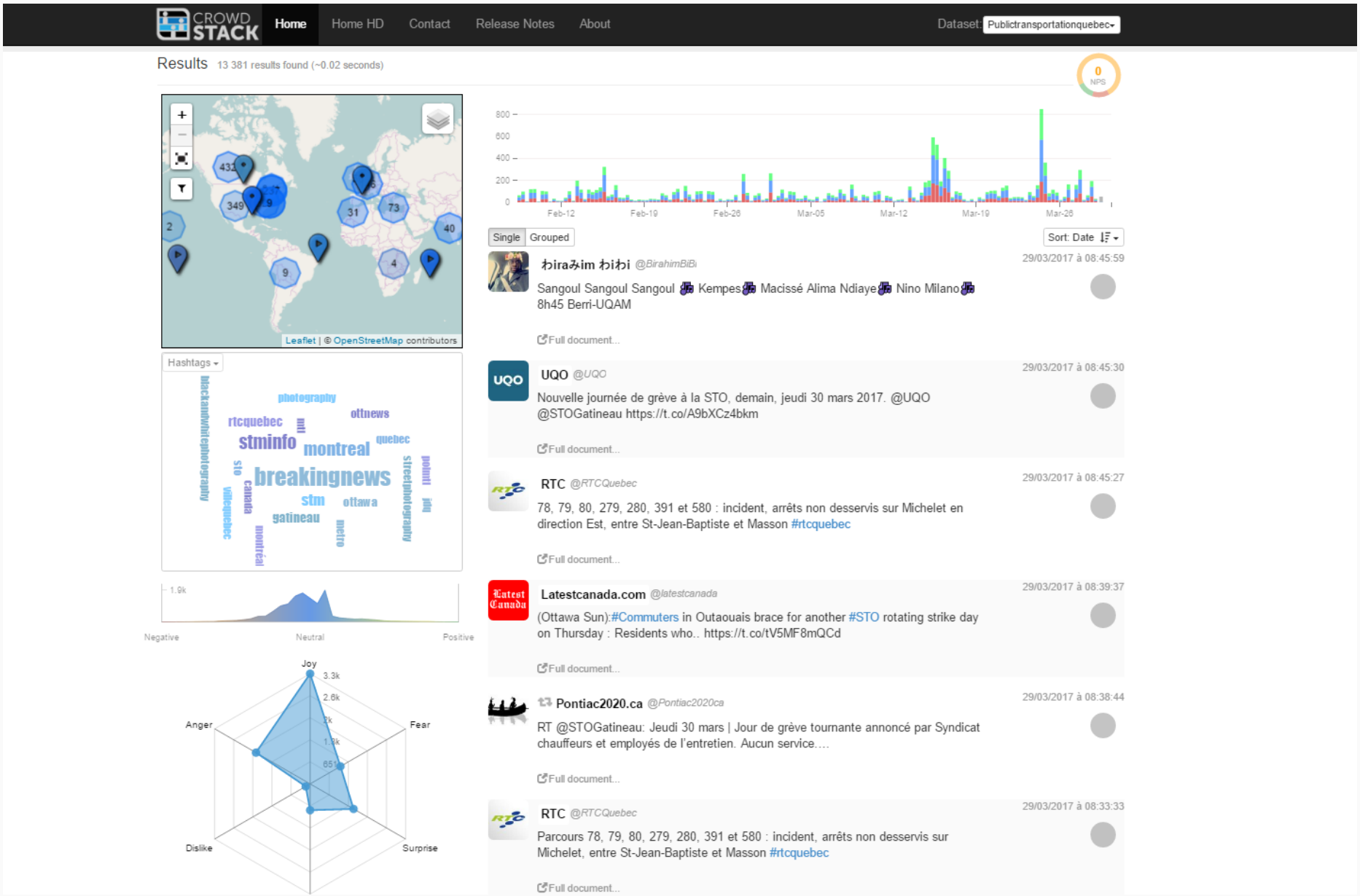
- direction : **Thierry Duchesne**, professeur au Département de mathématiques et de statistique
- codirection :

Christian Gagné : Professeur au Département de génie électrique et de génie informatique.

Via un stage Mitacs Accélération au sein de l'UMRsu.

Projet de Thales Canada - Crowdstack

Interface de Crowdstack



Objectif de ce partenariat

Ajouter des capacités d'analyse à Crowdstack

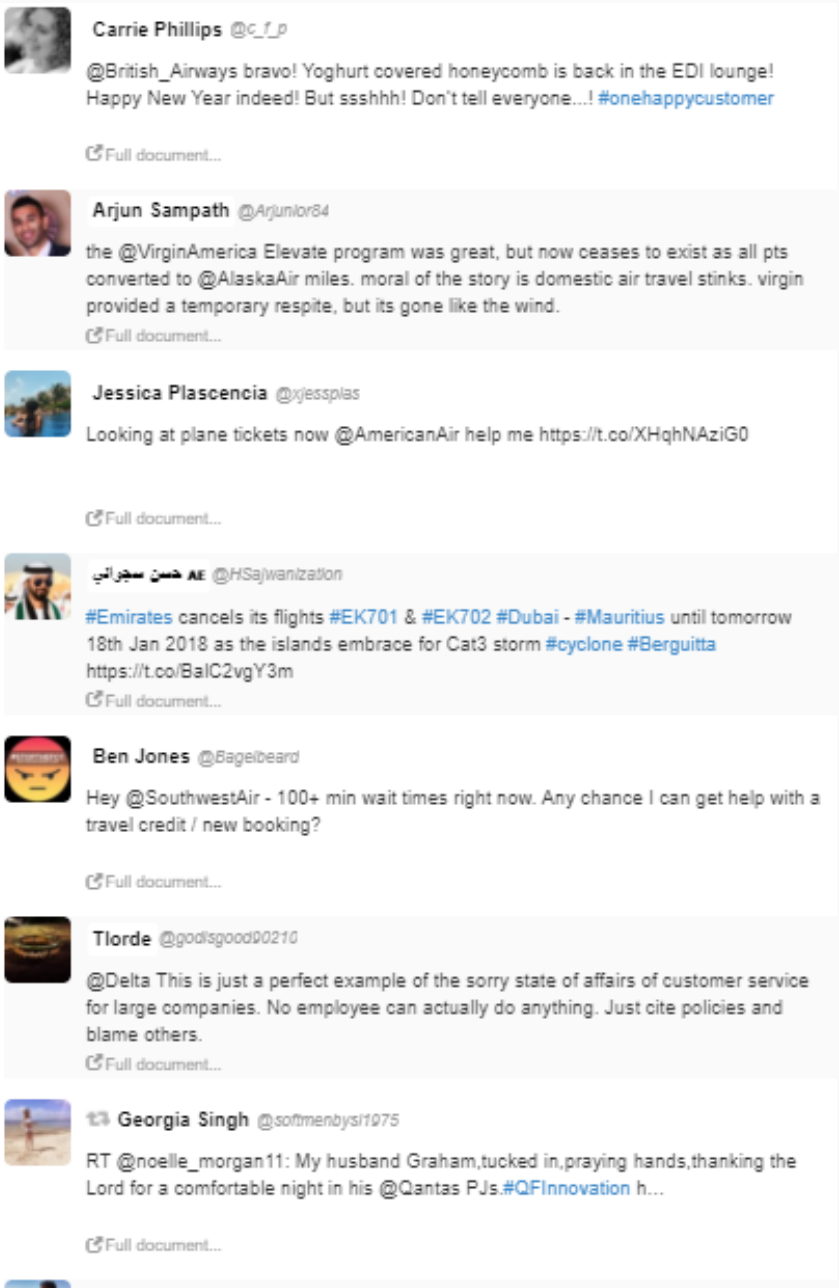
↓

Faciliter le travail d'un analyste humain qui souhaite :

- **extraire de l'information** à partir d'une banque de tweets passés ;
- faire de la **surveillance en temps réel** de flux de tweets, dans le but de détecter des événements.

Description des données

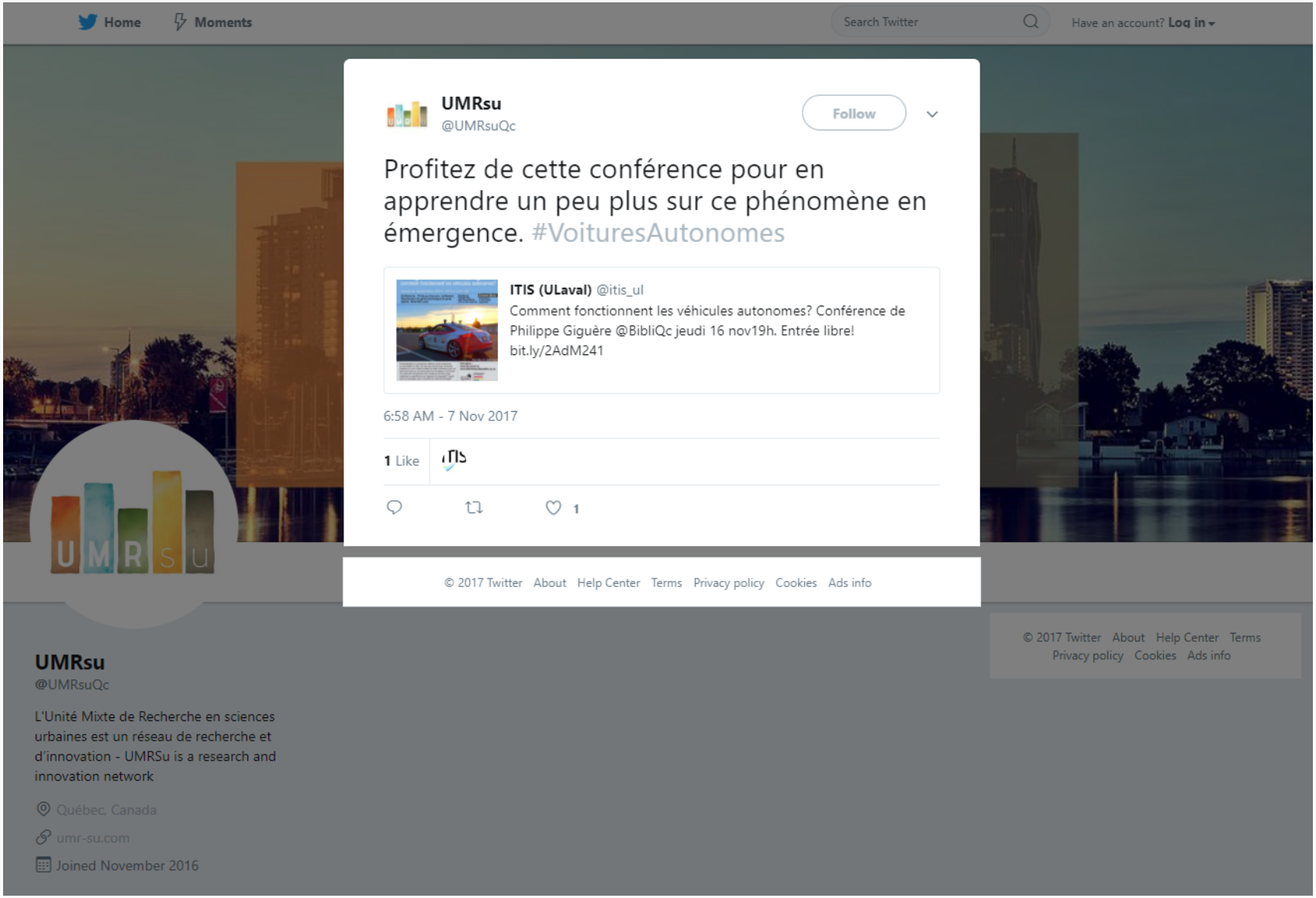
Exemple de flux de tweets



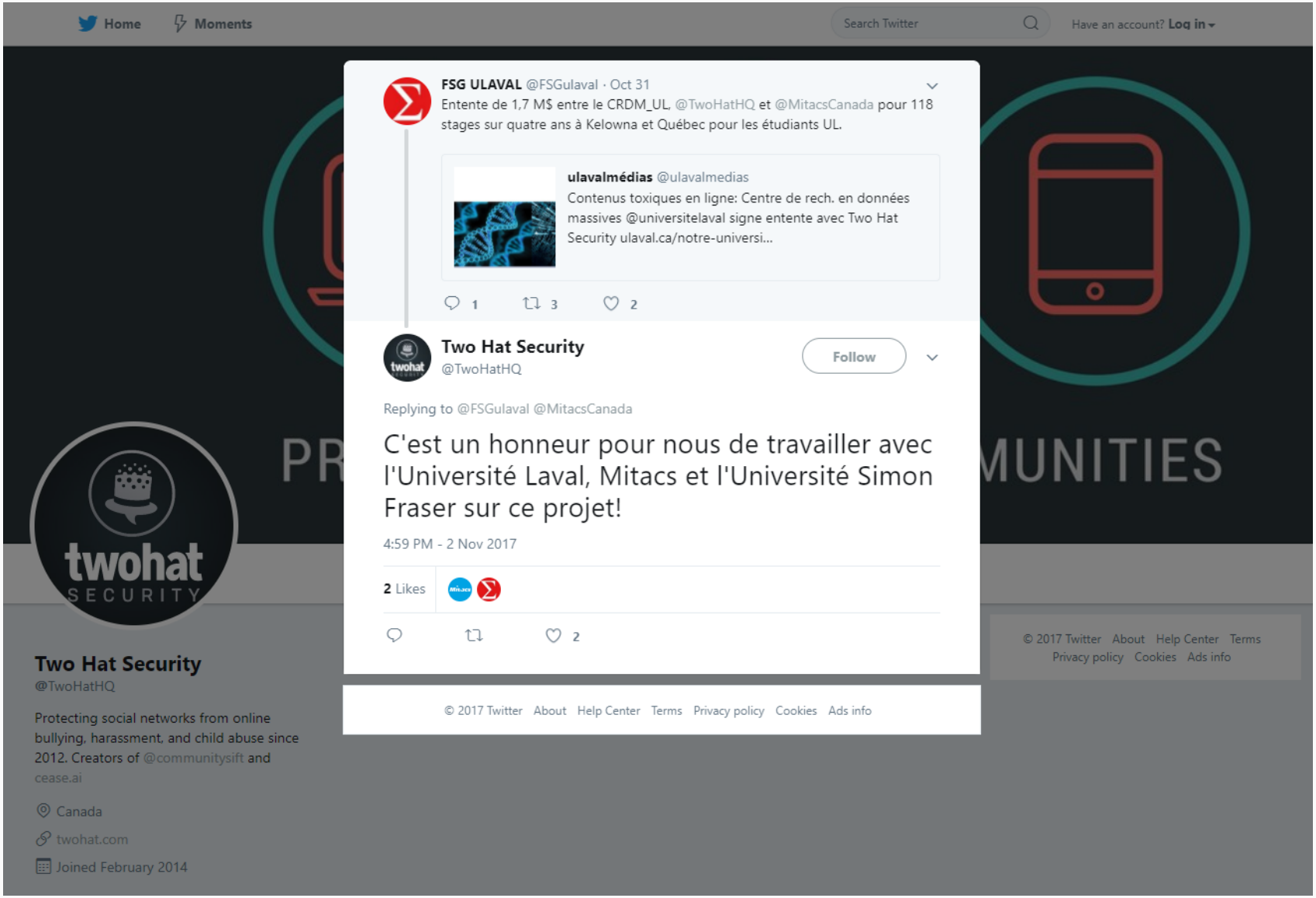
Exemple de tweet

Exemple de retweet

Exemple de tweet cité



Exemple de réponse à un tweet



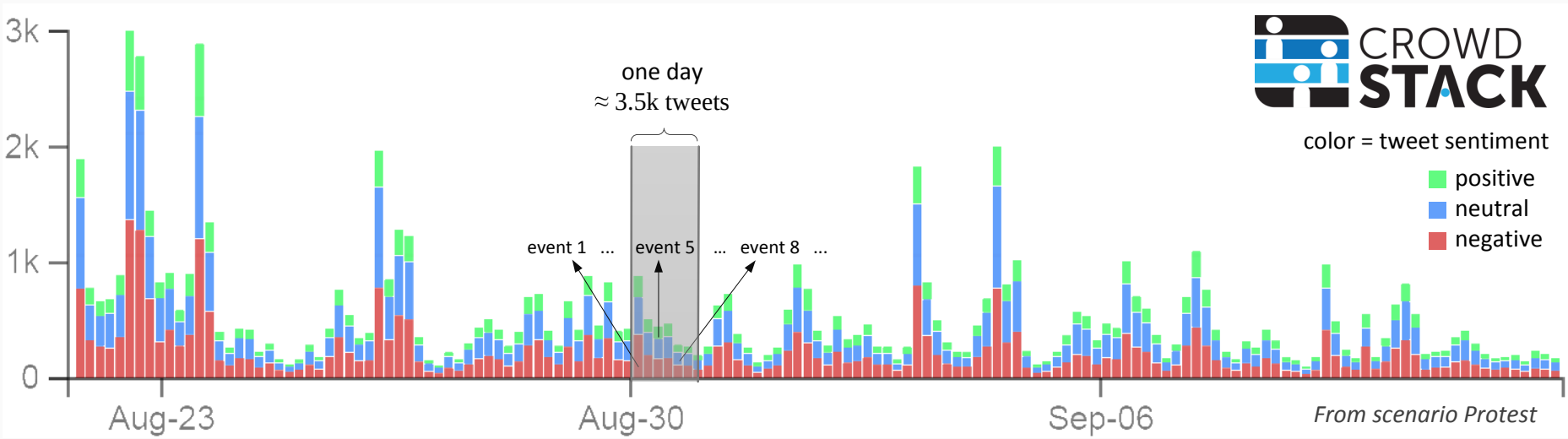
S'agit-il de données massives ?

Sommes-nous en présence de :

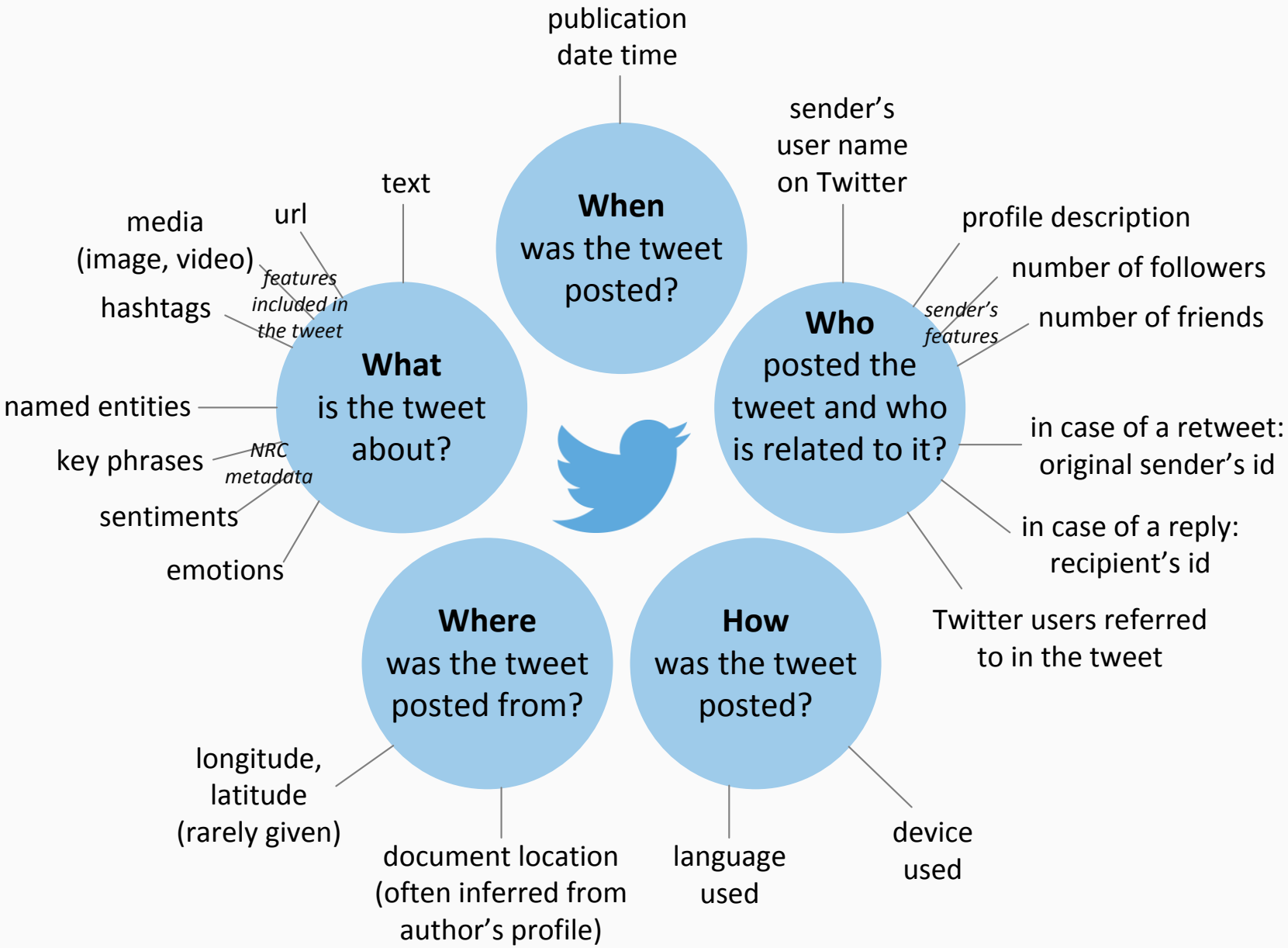
- **V**olume ?
- **V**élocité ?
- **V**ariété ?
- **V**éracité potentiellement douteuse ?

Volume et Vélocité

- Volume croissant, influencé par les filtres
- Vélocité parfois grande, variable dans le temps et entre les flux

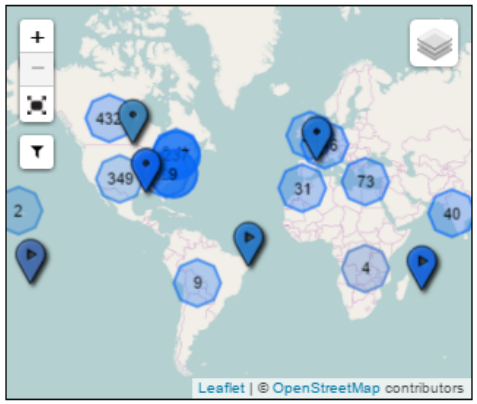


Variété : Un tweet, ce n'est pas que du texte !



Véracité : bruit et incertitude

- Bruit : plusieurs tweets non pertinents à filtrer
- Incertitude associée à des variables dérivées
  - localisation du tweet inférée à partir du profil de l'auteur



  - certaines métadonnées du CNRC proviennent de modèles
    - sentiments : probabilités d'être négatif, neutre ou positif.
    - émotions : scores de tristesse, joie, peur, aversion, colère, surprise.

Autres défis propres aux tweets

- redondance (retweets officiels, faux retweets)
- particularités du texte, par exemple :

> RT @Flarouche: Tjs impressionné par @RTCQuebec à chaque visite de #Qc. Écrans dans bus, marketing, propreté. @stminfo devrait s'en inspirer…

> Nothing new lol https://t.co/qoV2vnbbZV

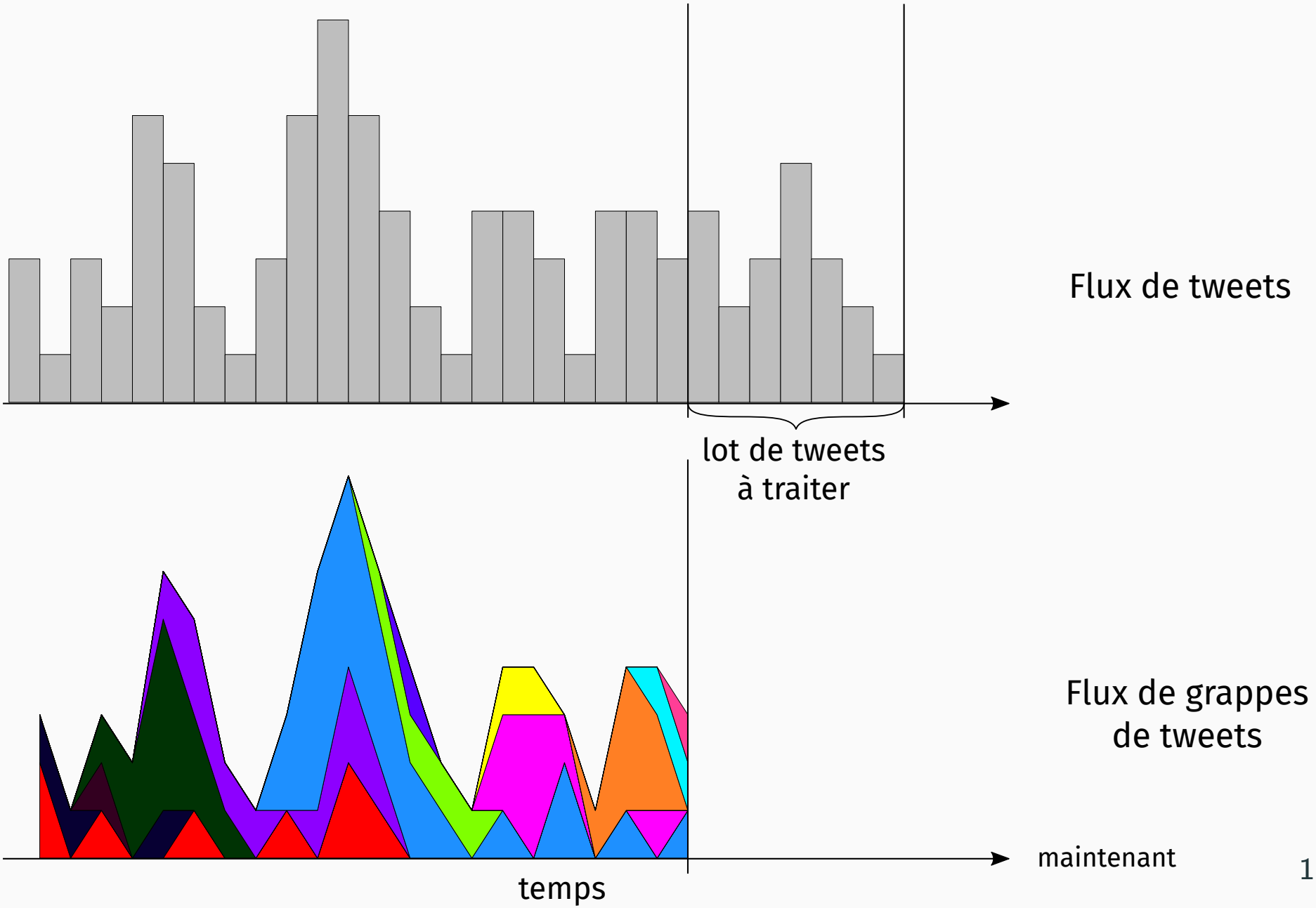
> Les midibus (800000$ chaque) du @RTCQuebec ont des prises USB !! Wow, on va pouvoir recharger nos Iphones 10-15min de charge 🤦‍♂️🤦‍♂️ !! #villequebec

Approche proposée

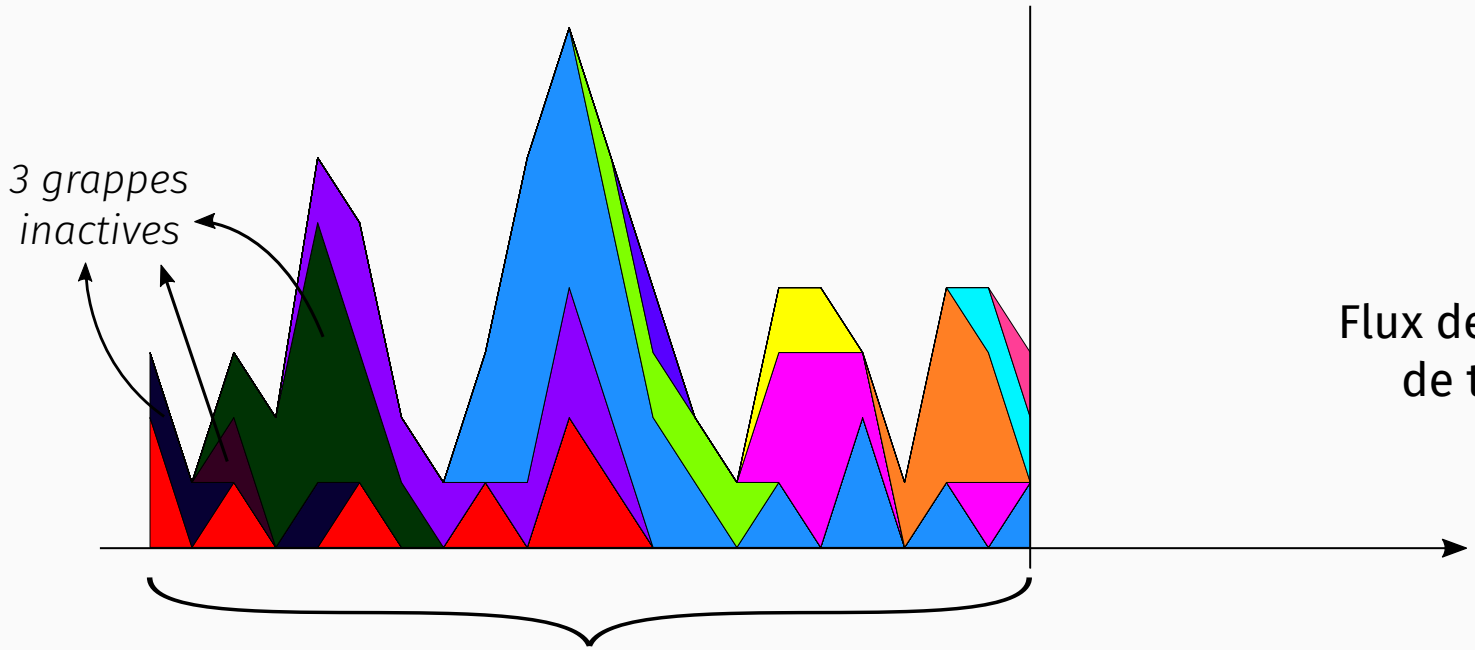
Fonctionnement de l'approche proposée

1. Regrouper les tweets similaires afin de réduire la quantité d'information à traiter.
2. Mettre en évidence les grappes de tweets intéressantes pour l'analyste.

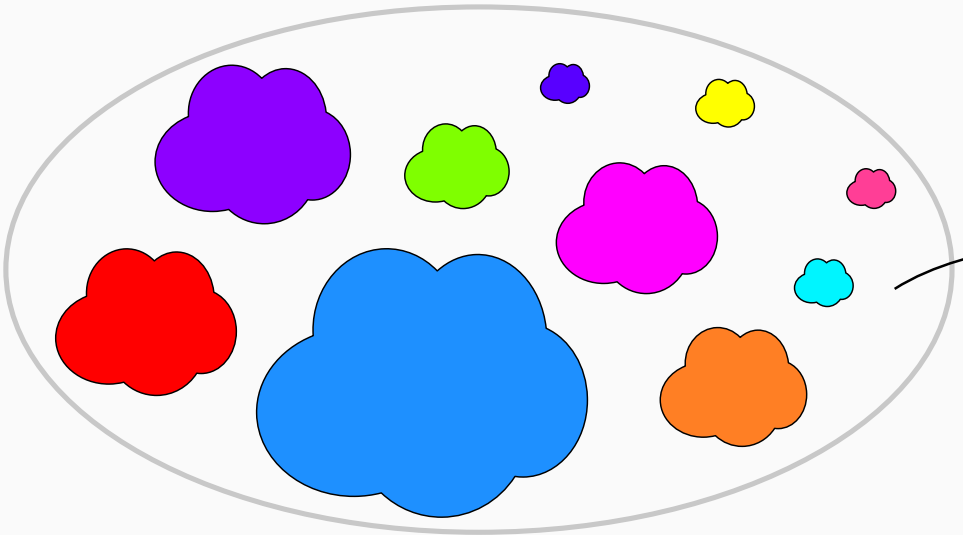
Données arrivant en temps réel



État des grappes à priori

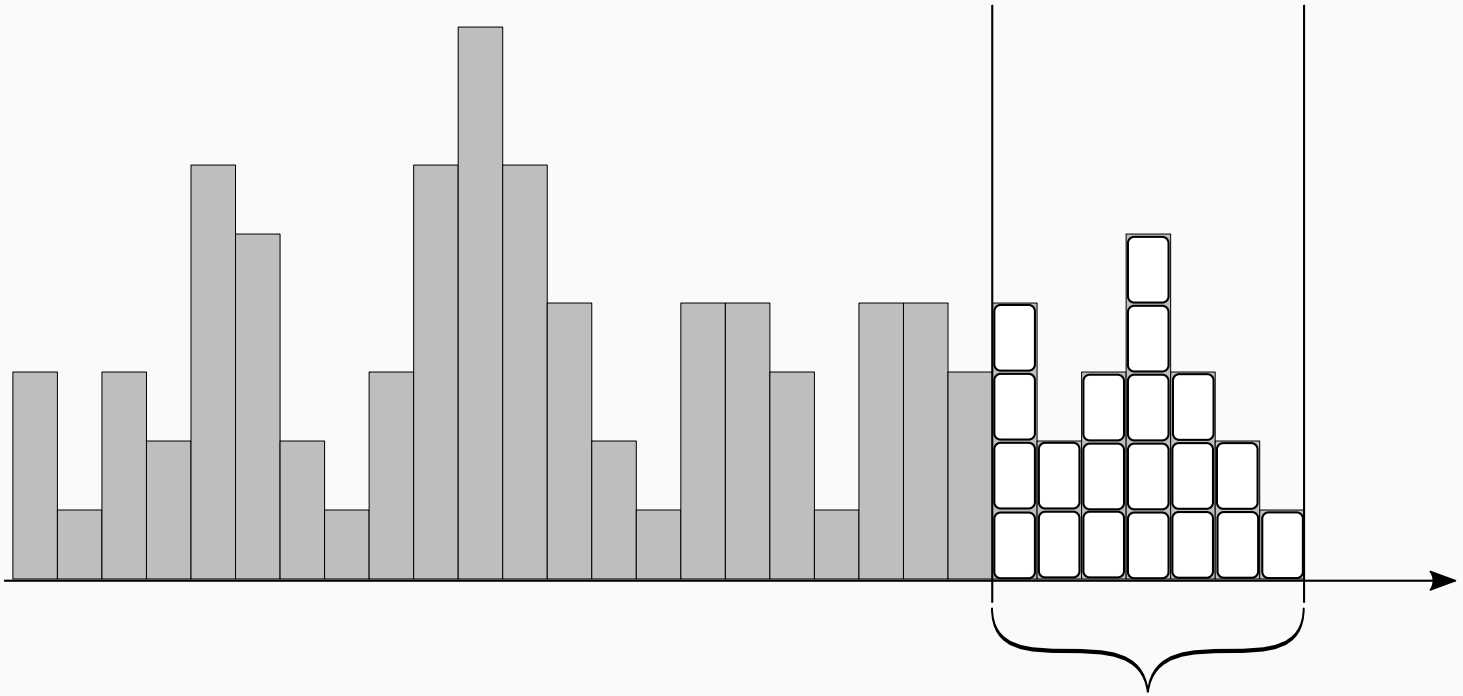


Flux de grappes
de tweets



grappes actives
pouvant potentiellement accueillir de nouveaux tweets

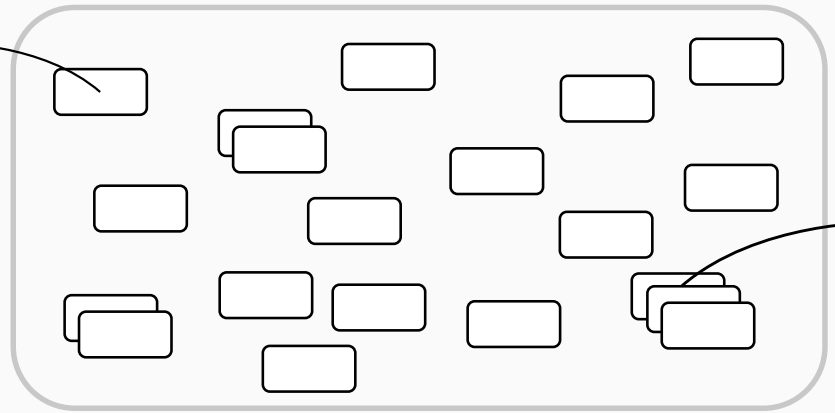
Traitement des données par petits-lots



Flux de tweets

Lot de tweets à traiter

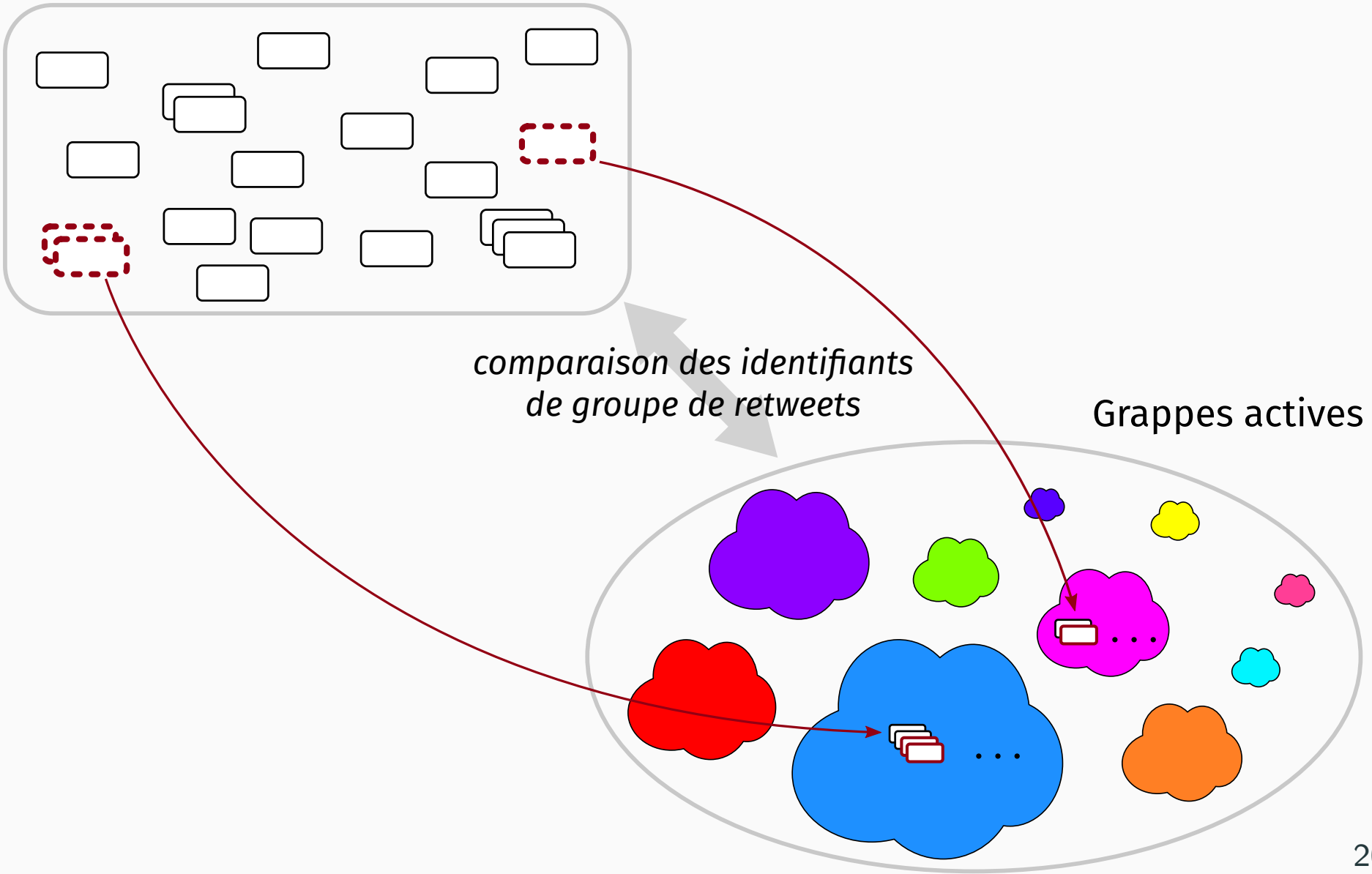
example de groupe de retweets comportant un seul tweet



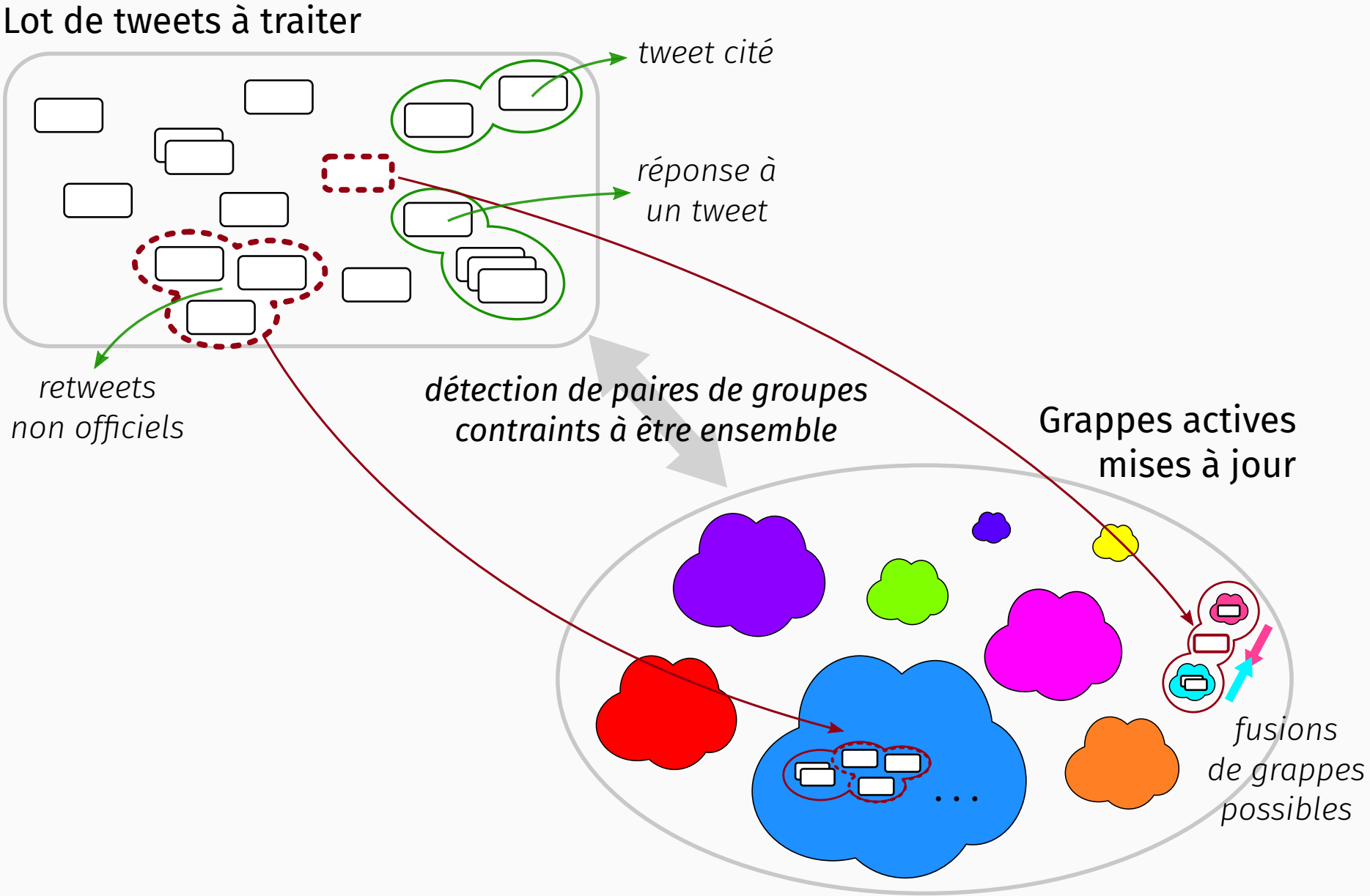
exemple de ***groupe de retweets*** *comportant 3 tweets*

Classement des groupes déjà vus

Lot de tweets à traiter

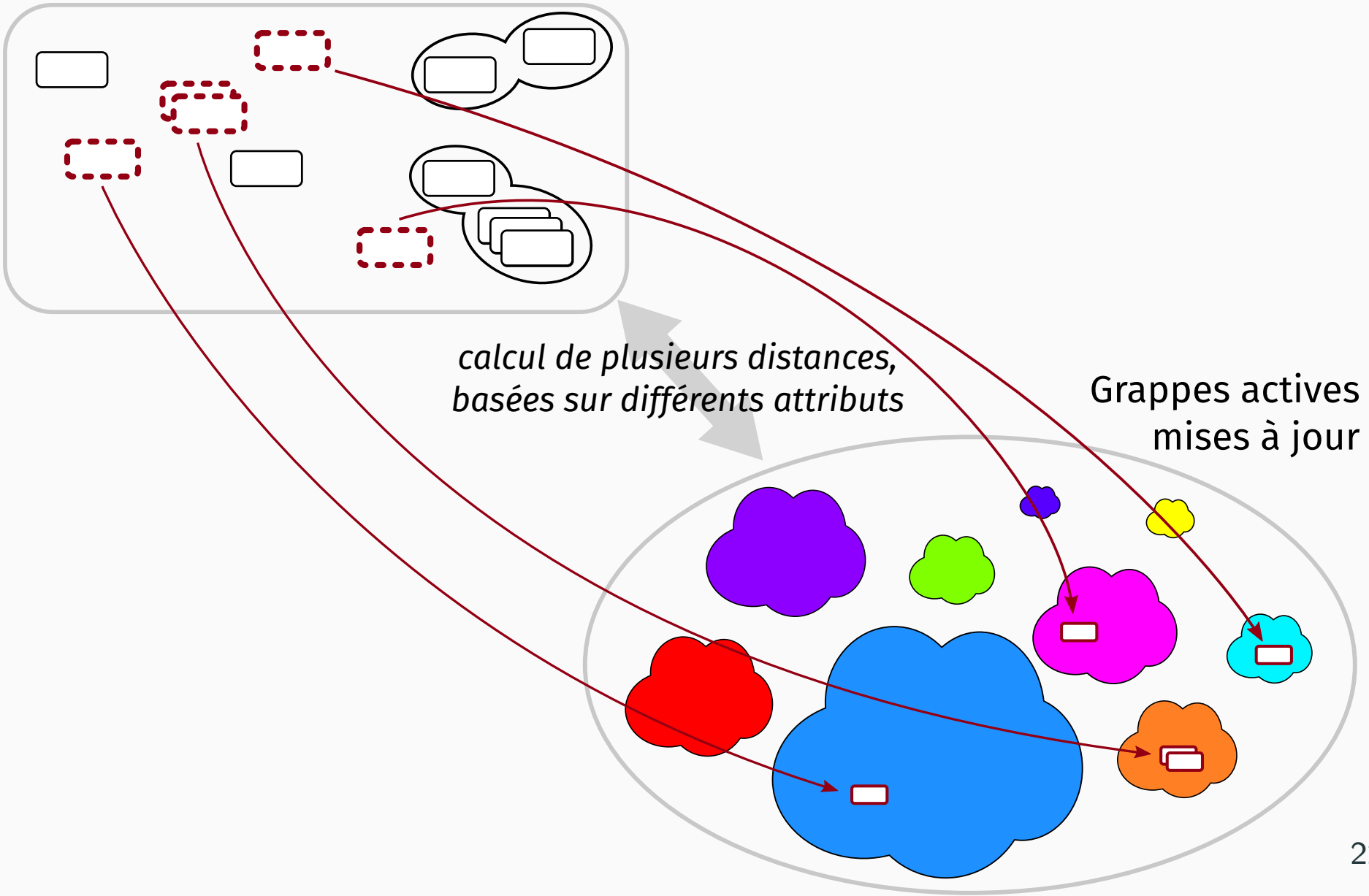


Classement par contrainte



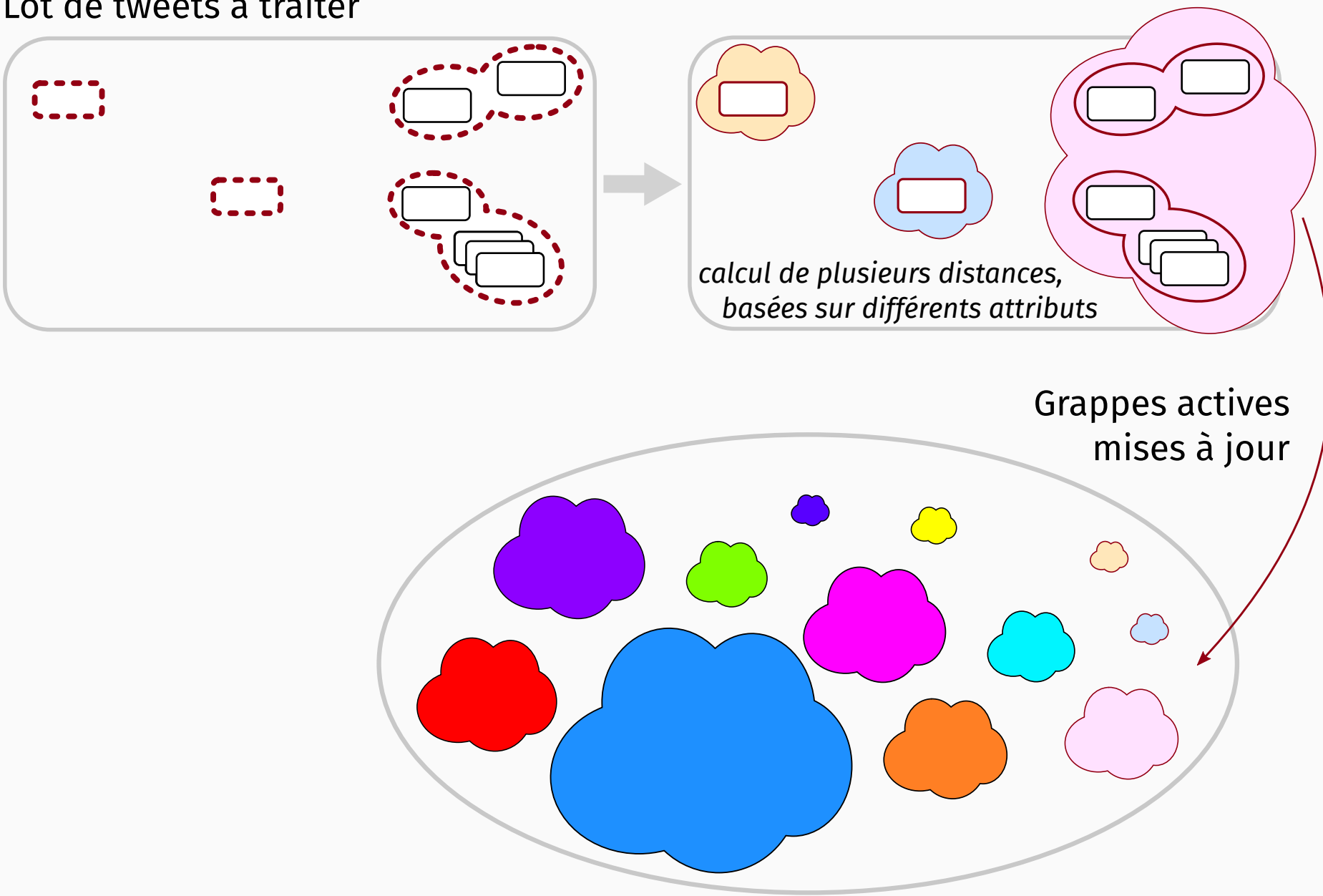
Classement non supervisé

Lot de tweets à traiter



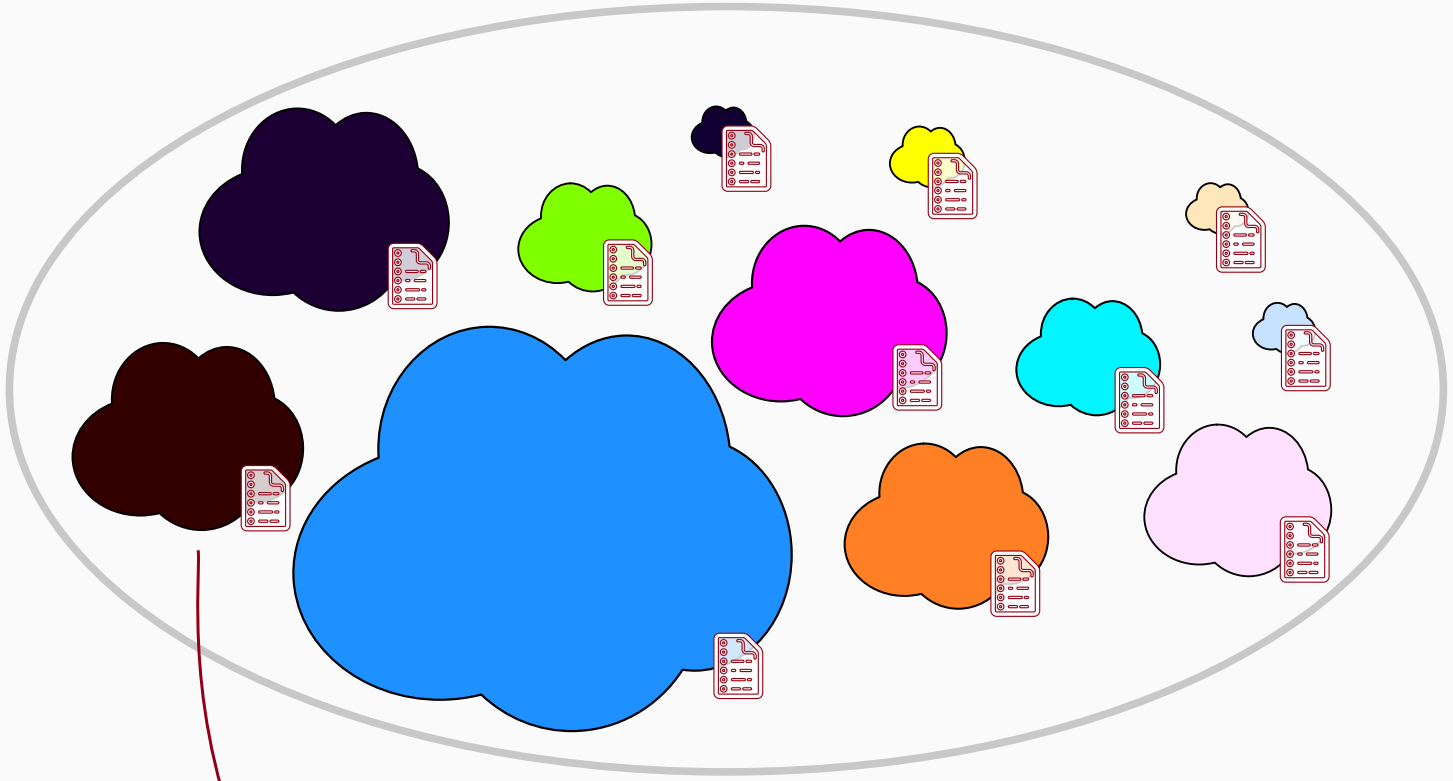
Formation de nouvelles grappes (clustering hiérarchique)

Lot de tweets à traiter



Qualification des grappes

Grappes actives

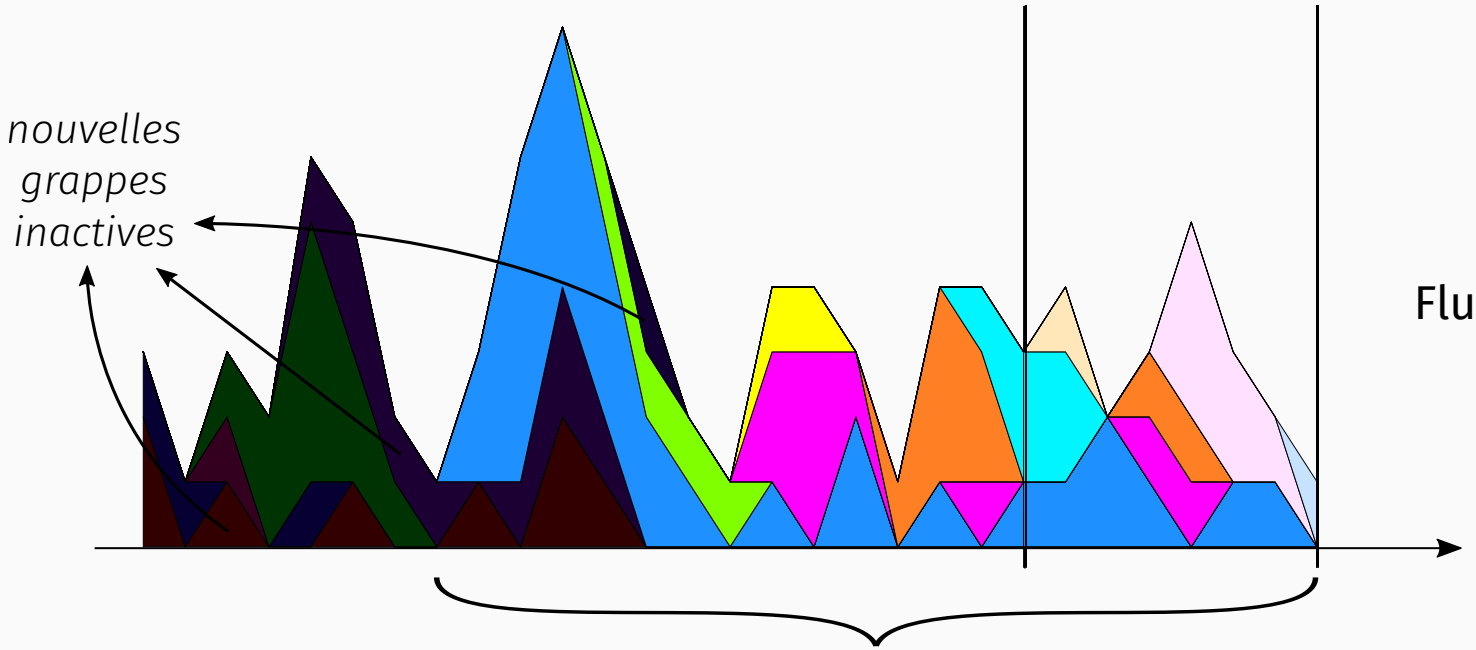


Des grappes peuvent être déclarées inactives.

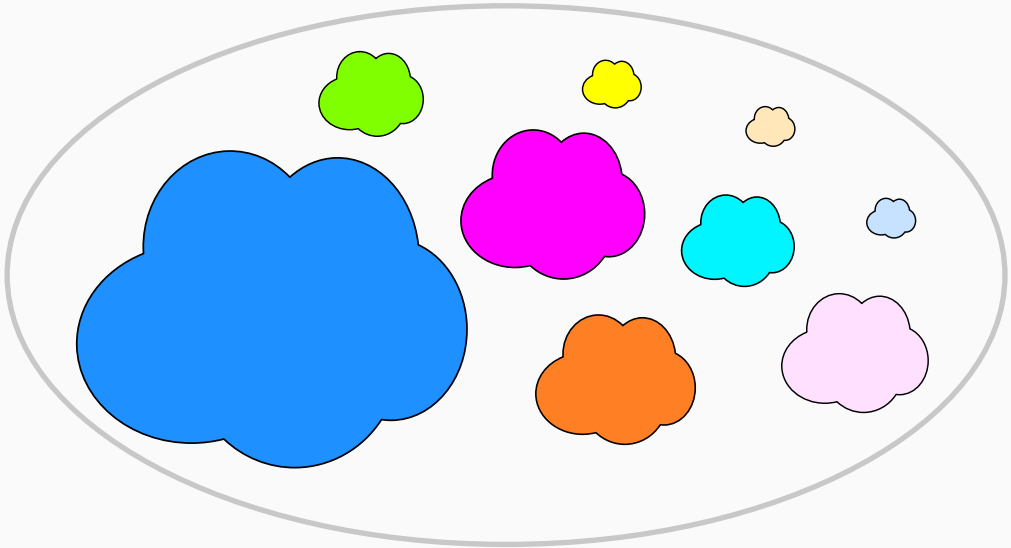
Mise à jour de :

- *score d'intérêt*
- *modèle d'activité*
- *statistiques agrégées (si la grappe a accueilli de nouveaux tweets)*

État des grappes à postériori



Flux de grappes de tweets mis à jour



Grappes actives mises à jour

Morceaux statistiques de l'approche

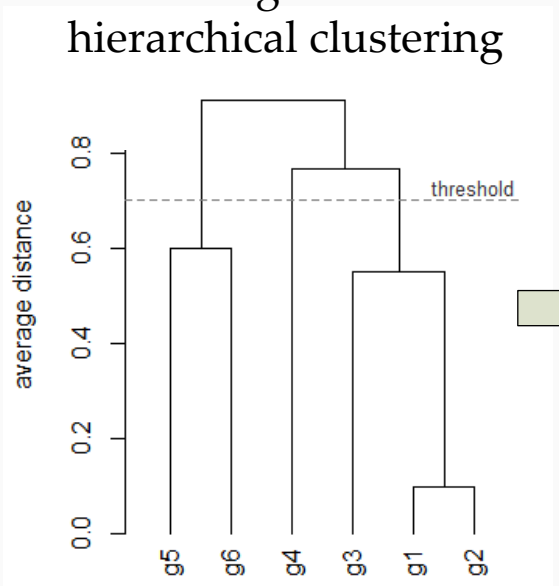
Clustering hiérarchique

Pour chaque distance considérée, basée sur une des nombreuses variables :

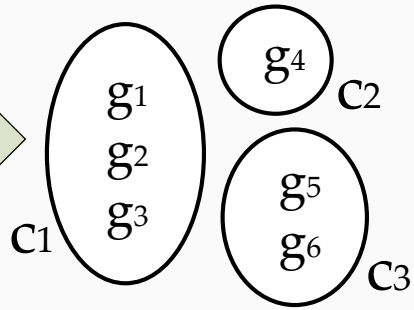
Group-group distance matrix

	g_1	g_2	g_3	g_4	g_5	g_6
g_1	0.0	0.1	0.7	0.9	1.0	0.8
g_2	0.1	0.0	0.4	0.6	0.9	1.0
g_3	0.7	0.4	0.0	0.8	1.0	0.7
g_4	0.9	0.6	0.8	0.0	1.0	0.9
g_5	1.0	0.9	1.0	1.0	0.0	0.6
g_6	0.8	1.0	0.7	0.9	0.6	0.0

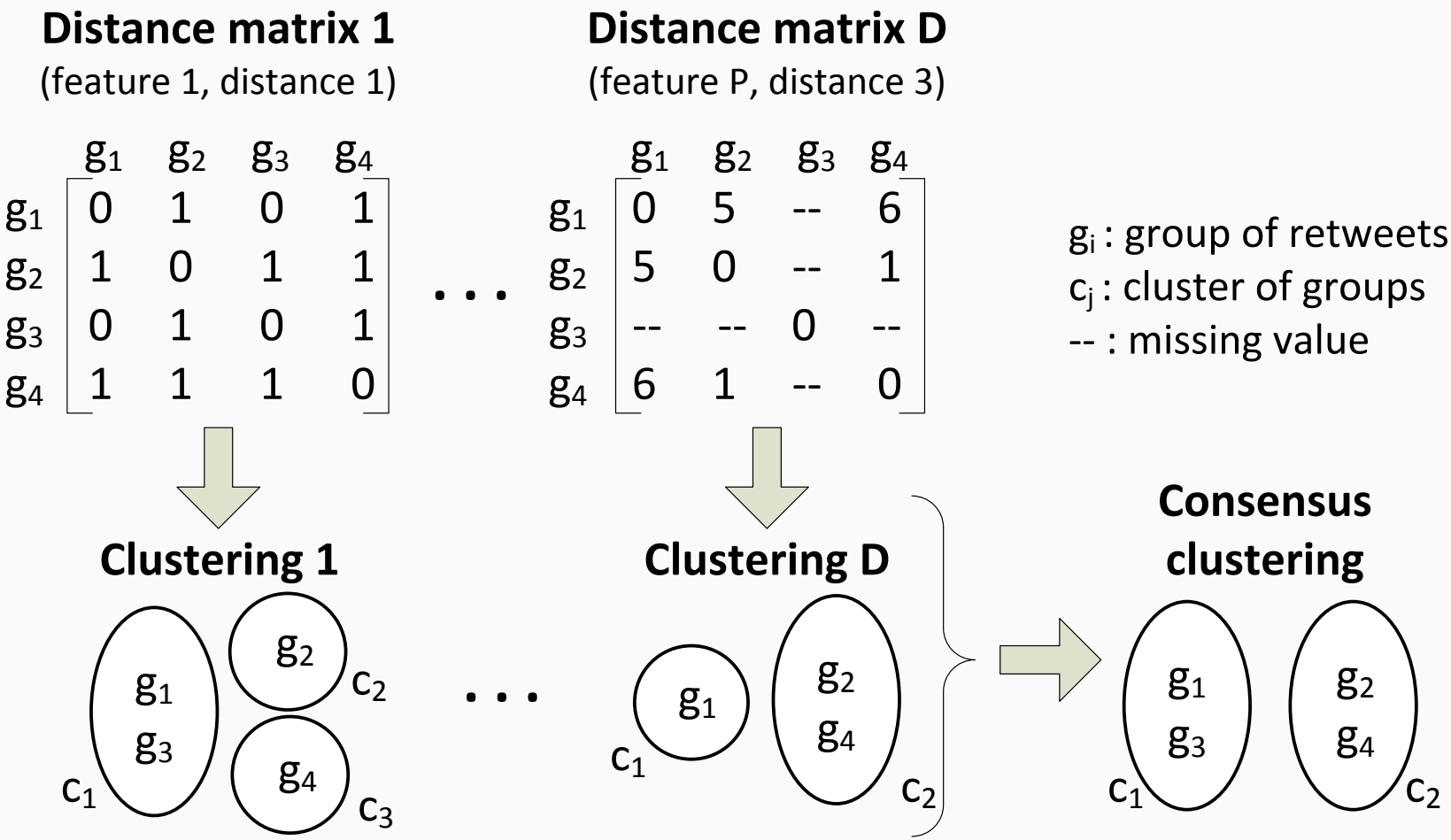
Dendrogram from hierarchical clustering



Partition after cutting the dendrogram

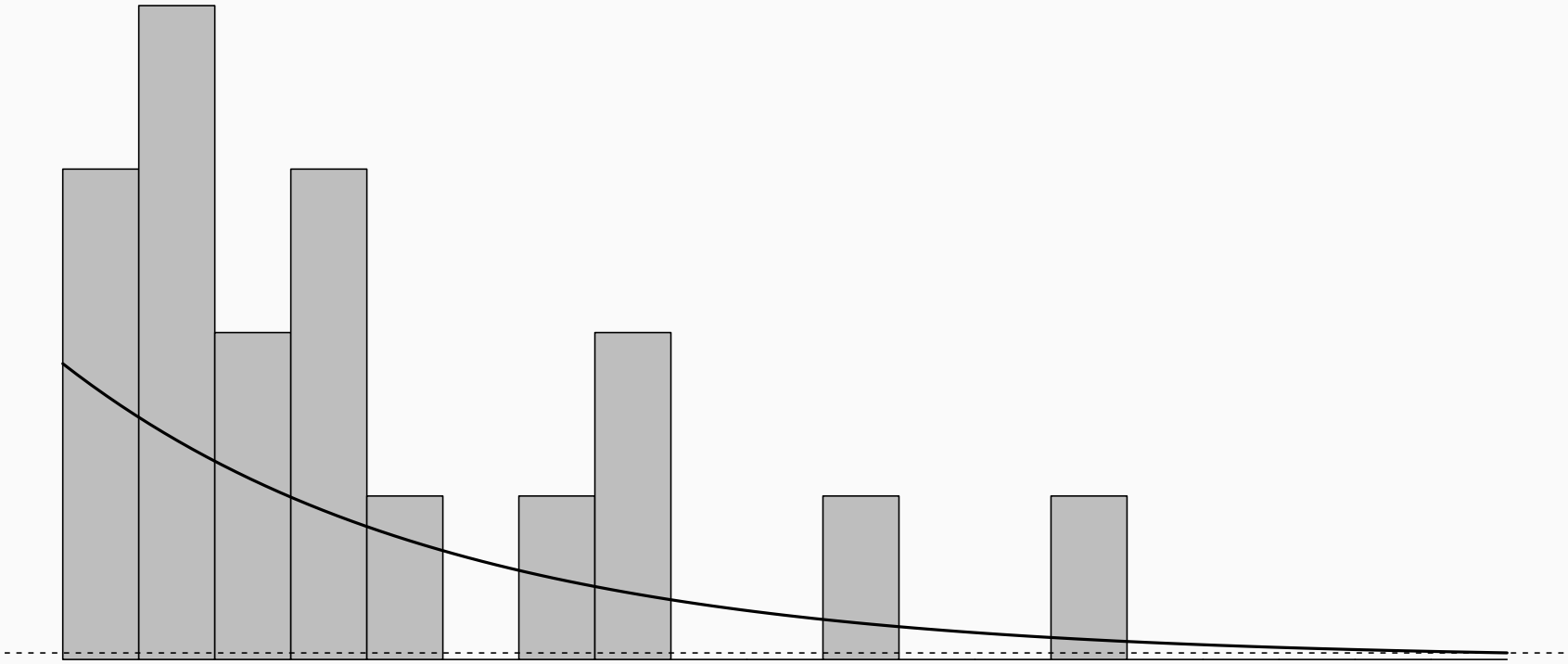


Clustering par ensemble



Modèle d'activité par grappe

Modélisation de l'arrivée des tweets dans la grappe :



Travaux futurs

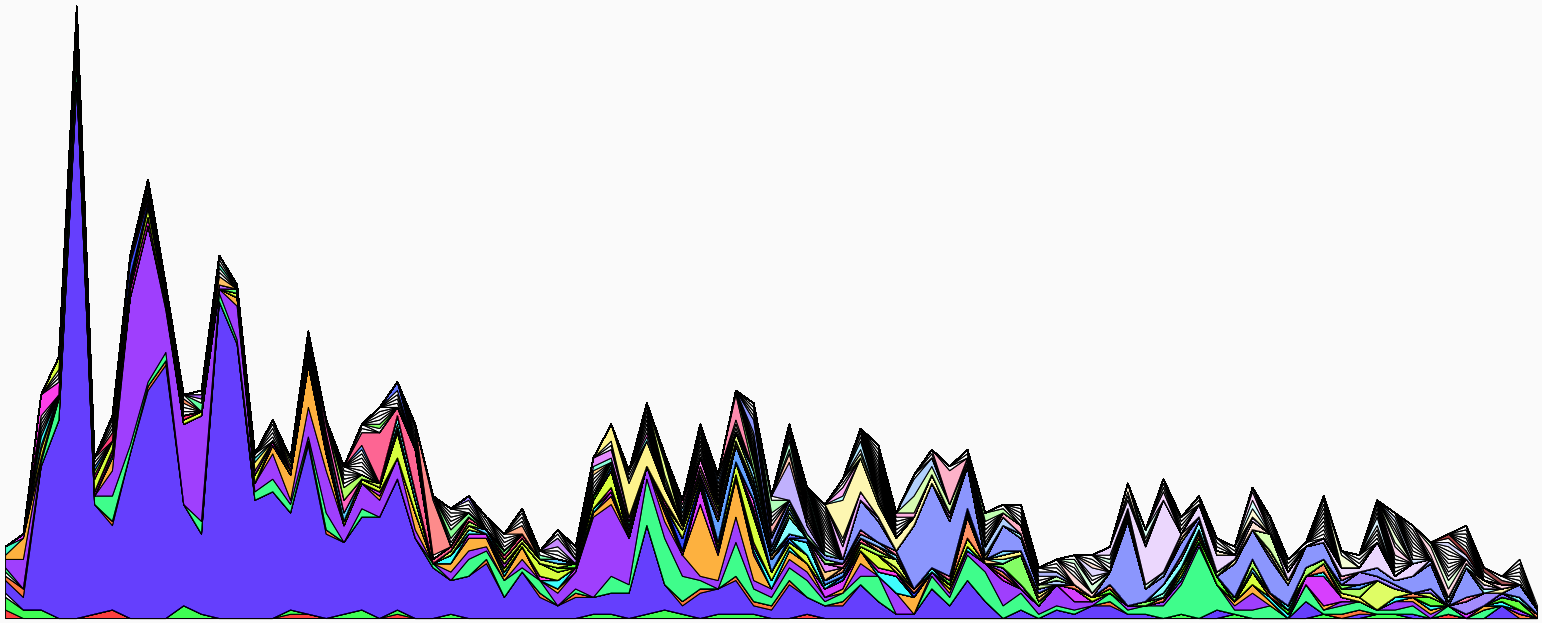
Travaux futurs

- Poursuivre l'implémentation de l'approche

elasticsearch

python

- Tester l'approche sur différents flux de tweets



- Publier l'approche

Merci !